RICE UNIVERSITY

Optimal Control of Flow and Transport Equations Using Discontinuous Galerkin Methods

by

Brianna Lynn

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Master of Arts

Approved, Thesis Committee:

Beatrice Riviere
Noah Harding Chair and Professor of
Computational and Applied Mathematics

Matthias Heinkenschloss
Professor of Computational and Applied
Mathematics

Adrianna Gillman
Assistant Professor of Computational and
Applied Mathematics

Houston, Texas

May, 2016

ABSTRACT

Optimal Control of Flow and Transport Equations Using Discontinuous Galerkin Methods

by

Brianna Lynn

This thesis analyzes the accuracy of discontinuous Galerkin methods for solving optimal control problems for flow and transport equations. The optimality conditions for each optimal control problem and error estimates for an optimal control problem constrained by a system of steady-state partial differential equations are derived. Synthetic data is used to create numerical examples that verify the methods work. Then the optimality conditions for the optimal control of the miscible displacement equations, where the control is the flow rate of the injection fluid, are derived.

Contents

Illustrations

Acknowledgements

I would like to thank my family for their love and support, my advisor, Dr. Riviere, for her encouragement and wisdom through the challenges of graduate school, Dr. Heinkenschoss, for improving my background in optimization, Dr. Gillman for enhancing my applied mathematics knowledge, and my undergraduate advisor and mentor, Dr. Anderson, for helping me develop a strong mathematical background.

Chapter 1

Literature Review

Though it is known that discontinuous Galerkin methods are well-suited for solving miscible displacement equations, less is known about the accuracy of these methods when used for optimal control problems governed by the miscible displacement equations. In this chapter, we will give a brief overview of research in topics related to this thesis. First, we will give background on the miscible displacement equations. Next, we will discuss research on optimal control problems governed by partial differential equations (PDEs). The last section will focus on optimization of reservoir flows.

1.1 Miscible displacement

The miscible displacement equations model the process of injecting a fluid into a reservoir to increase production of oil. These equations form a coupled system of a flow equation and a transport equation, where the unknown variables are the concentration and the pressure of the fluid mixture. These equations have been discretized by various methods, including classical methods (finite element methods, mixed methods, and finite volume methods) and discontinuous Galerkin methods.

Feng [9] proves existence and uniqueness of a solution to the miscible displacement problem in two dimensions. The paper proves the uniqueness of the semiclassical solution for the problem and also proves that the weak solution to the problem exists using the method of regularization.

1.1.1 Classical numerical methods

Ewing and Wheeler [8] derive a Galerkin method for solving the miscible displacement equations. The equations are discretized using finite element methods in space and a backward differencing method in time. *A priori* error estimates are then derived for the continuous in time problem, as well as for each discrete time step. The time stepping discretization is of first order.

Ewing and Russell [7] expand on the work in [8] by solving a more general problem. In the time discretization, the time step taken when solving for the pressure is larger than the time step taken for the concentration. The paper describes and then analyzes an iterative method to solve the miscible displacement problem at each time step. *A priori* error estimates are derived for solving for the concentration and pressure using this iterative method.

Douglas, Ewing and Wheeler [5] introduce a time stepping method for solving the miscible displacement problem. The pressure and the velocity are discretized in space using a mixed finite element method, while the concentration is discretized using a finite element method.

Russell [24] solves the miscible displacement problem using a sequential backward-difference time-stepping scheme. The pressure is approximated using a standard Galerkin method and the concentration is approximated using both a Galerkin method and a method of characteristics. The numerical scheme is explicitly stated and the error estimates are derived. The error estimates are also obtained for the problem with dispersion, though there is an iterative stabilization procedure added to the method.

Ohlberger [21] proposes a mixed finite element and finite volume method for solving a model problem with miscible and immiscible two phase flow. This discretization eliminates instability of standard finite element methods when the problem is convection dominated. The mixed finite element scheme is used for solving for the Darcy

velocity and the pressure, and the finite volume scheme is used for solving for the concentration. The convergence of the semi discrete scheme is shown to be of order one in space, with dependency on the diffusion coefficient. The convergence for the fully discrete scheme is proven to be order one in both space and time, also with dependency on the diffusion coefficient. A numerical example with synthetic data is given that validates the expected convergence rates.

Chen and Ewing [3] analyze the miscible displacement problem by proving the weak formulation for the problem converges to the exact solution as the mesh size decreases when using a finite element method. The paper then expands this process to the two phase flow and transport problem by proving the weak formulation for the problem also converges to the exact solution as the mesh size decreases when using a finite element method. Both of these problems can be used when modeling fluids in porous media for reservoir flows.

1.1.2 Discontinuous Galerkin methods

Riviere and Wheeler [23] propose an algorithm for solving the miscible displacement equations using discontinuous Galerkin (DG) methods in space and the backward Euler method in time. Numerical results are given when solving the problem on structured and unstructured meshes. These results show that DG methods work well for solving fluid flow problems, are competitive with other methods, and give the ability to use unstructured meshes easily.

Sun et al. [27] introduce a mixed finite element method and discontinuous Galerkin method for solving the miscible displacement equations. The flow equation is discretized using a mixed finite element method and the transport equation is discretized using a discontinuous Galerkin method. Error estimates are proved for the flow equation and transport equation separately, as well as for the coupled system. The authors

give parameter choices for the discretization that will produce the optimal convergence rates for the concentration and the velocity.

Epshteyn and Riviere [6] study the miscible displacement equations using primal discontinuous Galerkin methods in space and backward Euler in time. The existence and convergence of the numerical solution are proven and the parameters for which the method is stable and convergent are given.

Bartels, Jensen and Muller [1] analyze a mixed finite element combined with a discontinuous Galerkin method when solving the miscible displacement equations under low regularity. An algorithm is presented that solves for the pressure, concentration and velocity given an initial concentration. The solutions determined from this algorithm are shown to exist and be bounded. It is proven that if the weak solution is unique, the numerical solutions of the pressure, concentration and velocity converge to a unique triplet that satisfies the weak formulation of the problem as the mesh size and time step size approach zero. Numerical examples are given to validate the accuracy of the algorithm and method, which is first order in time.

Riviere and Walkington [22] analyze the convergence of solving the miscible displacement equations using mixed finite element and continuous finite element methods in space and discontinuous Galerkin methods in time. This discretization gives high order accuracy in time as well as convergence under low regularity. Some of the assumptions needed for convergence using other methods can be relaxed while still ensuring stability and convergence for this method. The diffusion tensor may be unbounded and various coefficients may be discontinuous. It is proven that the numerical solutions for the velocity, concentration and pressure converge to a weak solution of the PDE. A high order in time scheme is derived and the scheme is proven to be stable.

Li [13] presents a numerical method for solving the miscible displacement equa-

tions using a mixed method in space and discontinuous Galerkin in time for the transport equation. He proves consistency and stability of the scheme and gives numerical results in 2D using the software DUNE. The method is flexible with high accuracy, which can be seen in the convergence rates of the concentration and pressure. The convergence rates increase as the order of the DG in time increases. Li also gives numerical examples of two physical problems: homogeneous grain size and homogeneous grain size with a discontinuous lens. The methods show high accuracy for these problems, but for the homogeneous grain size problem, there is an overshoot and undershoot seen when solving the transport equation with higher order DG methods.

Li et al. [15] solve the miscible displacement equations using a mixed discontinuous Galerkin method in space and discontinuous Galerkin in time. The method is high order in time and space for the solution of the miscible displacement equations under low regularity. A general Aubin-Lions theorem, which is valid for broken Sobolev spaces, is proven.

1.2 Optimal control problems with PDE constraints

Optimal control problems are a specific type of optimization problem where the goal is to minimize an objective function, or cost function, that is subject to certain constraints. The problems we focus on have PDE constraints, which means that the goal is to minimize a function $J(y(u), u)$ where y must solve a given PDE or system of PDEs that depend on u. The technique is to minimize J over u in U, which is the control space, and the PDE is the solved for y in Y, the state space, given u.

To minimize the objective function, standard optimization techniques can be used (see Section 1.2.1). The PDE can be solved using many different numerical methods in space, including finite element methods, finite difference methods, and discontinu-

ous Galerkin methods and Runge-Kutta methods in time. For more information on optimal control problems with PDE constraints, see [10], [16] and [28].

1.2.1 Optimization

There are many optimization methods that can be used to minimize the objective function $J(u)$ in an optimal control problem. Some of the more common methods are Newton methods, line search methods, trust-region methods and conjugate gradient methods, which are all iterative methods. A few examples of these optimization methods are presented here. More detail is given in [20].

Newton's method requires the Hessian and the Jacobian of the objective function J. The iterative step comes from solving for the Newton step, v_k, defined by

$$\nabla^2 J(u_k)v_k = -\nabla J(u_k).$$

The next iteration is defined as

$$u_{k+1} = u_k + v_k.$$

Newton's method works well if the initial guess u_0 is close enough to the exact solution u_*. Since the exact solution is unknown, it can be difficult to pick a u_0 such that it lives in the basin of attraction for u_*. But if the Hessian of J is uniformly positive definite, then Newton's method will converge globally, for any initial guess.

Line search methods compute a search direction p_k and then determine the distance α_k to go in that direction. Given u_k, the next iteration is given by

$$u_{k+1} = u_k + \alpha_k p_k.$$

To guarantee that the value of the objective function decreases as one goes along the direction p_k with sufficiently small step size α_k, p_k needs to be a descent direction, which must satisfy

$$p_k^T \nabla J(u_k) \leq 0.$$

The search direction is frequently defined using B_k, a symmetric and nonsingular matrix, by

$$p_k = -B_k^{-1}\nabla J(u_k).$$

Two common methods that use the search direction in this form are the steepest descent method, where B_k is the identity matrix, and Newton's method, where B_k is the Hessian of the objective function evaluated at u_k. There are different ways to determine the step length α_k, including the Wolfe conditions and the Goldstein conditions, as well as an Armijo backtracking line search. (See [20] for more information on these conditions.) The convergence of line search methods depends on the search direction and the step length.

Trust-regions methods pick a region that is an adequate representation of the objective function at u_k, then pick a step length and search direction within this region. If $J(u_{k+1})$ is greater than or equal to $J(u_k)$, the method shrinks the size of the trust region and then determines the next iteration.

Conjugate gradient methods are very useful for convex quadratic optimization problems because they can be used to solve large linear systems and nonlinear optimization problems. Similarly to the previous methods, conjugate gradient methods determine a search direction, called a conjugate direction, and a step size. Let H be an $n \times n$, symmetric positive definitive matrix and consider a problem of the form

$$Hu = b.$$

The conjugate directions, $\{p_0, p_1, ..., p_{n-1}\}$, are defined from the matrix H such that

$$p_i^T H p_j = 0, \quad \forall\, i \neq j.$$

The step size is defined by α_k:

$$\alpha_k = -\frac{(Hu_k - b)^T p_k}{p_k^T H p_k}.$$

Thus we have our next iteration, defined by

$$u_{k+1} = u_k + \alpha_k p_k.$$

One of the useful aspects of this method is the following theorem [20]: Given any initial guess u_0, the sequence $\{u_k\}_{k\geq 1}$ converges in at most n steps to the solution u_* of the system

$$Au = b.$$

For more information on optimization techniques, see [20].

1.2.2 Optimal control problems with PDE constraints discretized by discontinuous Galerkin methods

Meidner and Vexler [19] determine error estimates for solving optimal control problems governed by linear parabolic PDEs. After explaining the full discretization of the problem, as well as giving assumptions needed for uniqueness of the solution, stability estimates are given for the state and adjoint equations. The state equation is discretized using a discontinuous Galerkin method in time and a continuous finite element method in space. The control is discretized using a discontinuous Galerkin method in time and two different ways in space: a discontinuous Galerkin method and a continuous finite element method. Error estimates are then given for the discretization of the state and control. It is shown that the optimal error for the state and the control in this discretization can be determined and is of order $O(\Delta t + h^2)$, where Δt is the maximum time step size, and h is the maximum diameter of the elements. Numerical results then verify the error estimates.

Yucel et al. [29] study the discontinuous Galerkin methods for solving optimal control problems governed by linear steady state PDEs using two optimization approaches: *discretize-then-optimize (DO)* and *optimize-then-discretize (OD)*. Given k to be the degree of the polynomial basis functions and h the mesh size, the paper

shows that when using the symmetric interior penalty Galerkin (SIPG) method, the DO and OD methods produce the same linear system, with L^2 error of $O(h^{k+1})$ for both the control and the state. When using nonsymmetric interior penalty Galerkin (NIPG) or incomplete interior penalty Galerkin (IIPG) methods, the OD and DO methods do not produce the same system nor do they give the same errors. The NIPG method produces L^2 error of $O(h^k)$ for OD and $O(h)$ for DO. This paper gives background as well as convergence rates for linear PDE constrained optimal control problems using different DG methods.

Leykekhman [11] discusses commutativity of discontinuous Galerkin methods for solving optimal control problems with PDE constraints. The optimality conditions and both the DO and OD systems for three different PDE constraints, including the advection-diffusion-reaction equation, are derived. When discretizing the advection-diffusion-reaction equation using SIPG, the DO and OD systems are the same, which means SIPG is commutative. When discretizing the PDE using NIPG, the systems are different, which means that NIPG is not commutative. Then the error estimates for the advection-diffusion-reaction equation problem using SIPG are derived. It is proved that given mesh size h and polynomial basis function degree k, the energy error of the control is order $O(h^k)$. Numerical results are then presented that validate the error estimates.

Leykekhman and Heinkenschloss [12] analyze the DG SIPG method for the optimal control of advection-dominated elliptic PDEs. The paper focuses on the advection-diffusion-reaction equation, where the coefficient of the diffusion term is very small. Error estimates are derived for this problem for when there are interior layers and when there are boundary layers. Numerical results are then given to validate the error estimates. The results show that the energy error of the control is $O(h^k)$ and the L^2 error is $O(h^{k+1})$, where h is the mesh size and k is the polynomial degree.

In his technical report, Heinkenschloss [17] describes numerical algorithms for solving implicity constrained optimization problems. The report focuses on solving optimal control problems with nonlinear PDEs as constraints. Two algorithms are given for solving many different arbitrary optimal control problems with nonlinear PDE constraints. One algorithm is given for solving an optimal control problem is Newton-Conjugate Gradient Method with Armijo Line-Search. The other algorithms describe how to compute the gradient and the Hessian of the objective function, which can be complicated. Heinkenschloss uses the algorithms to solve the optimal control of the Burgers' equation, and then gives a numerical example to verify that the optimization method converges to the solution. Explanation of how to efficiently program the algorithms are given, including ways to avoid recomputing variables.

1.3 Optimization in oil recovery

Chavent and Dupuy [2] use optimal control theory and history matching of the pressure to determine the permeability distribution in a single phase flow problem. The PDEs are discretized using finite differences in space and Crank-Nicolson in time. To solve the optimization problem, numerical results are given that show stability and flexibility of the method. The authors state that some of the benefits of their method are speed of computation of individual cell values and avoidance of unrealistic parameter values.

Mehos and Ramirez [18] study the miscible displacement of oil caused by carbon dioxide flooding modeled using an optimal control problem. The function they maximize is the value of oil produced minus cost of carbon dioxide injected plus value of carbon dioxide recovered. The PDEs are approximated using finite differences, implicit in the pressure and explicit in the saturation, and the optimization is solved using an iterative gradient method. Numerical simulations are given for three differ-

ent cases: a single carbon dioxide slug, simultaneous injection of CO_2 and water, and water-alternating-gas injection. The numerical results show that for each case, there is roughly a 5.9 million dollar profit.

Zeitout and Pinder [30] solve a coupled flow and transport problem using an optimal control least squares approach. The optimization of this problem is solved using a new conjugate gradient algorithm after preconditioning the system. The authors used numerical results to compare two methods for solving this problem: classic finite element method and optimal control least squares method. Though the finite element method solved the problem faster, the method caused oscillations in the approximate solution, depending on the Peclet number. These oscillations did not occur in the least squares approach.

Simon and Ulbrich [25] study the optimal control of partially miscible two-phase flow, where the objective function models the amount of trapped CO_2 after injection into a reservoir. The control is the injection rate of the carbon dioxide. To discretize the PDEs, the authors use a BOX method, which is a locally conservative control-volume finite element method in space, and backward Euler in time. The package Sundance, which allows for unstructured grids and parallelization, is used to run the simulations of the problem. Some of their results include that two wells must be active for optimal results and that there is an upper bound on the amount of CO_2 injected into the well for optimal results.

Simon and Ulbrich [26] expand on their research in [25], on partially miscible two-phase flow problems. Solving the same problem as in the previous work, the authors use a periodic averaging feature to eliminate oscilliations caused by the spatial discretization. The algorithm presented is parallelizable, can be extended to more complicated reservoir models, and links to the state-of-the-art interior point optimization software.

Chapter 2

Optimal Control of a Steady-State System of Equations

This chapter describes the solution of an optimal control problem with linear PDE constraints, which form a system of steady-state elliptic equations. We will derive the continuous optimality conditions for the problem, which are independent of the discretization used. We will then go through the discretization in space using discontinuous Galerkin methods. We prove error estimates for the discrete optimal control problem. Next, we will derive the discrete optimality conditions using the Lagrangian of the problem. We will then describe the optimization algorithm to solve the problem. Last, we will give numerical results to validate the methods chosen.

2.1 Problem statement

The goal of this work is to solve an optimal control problem governed by a system of linear, steady-state PDEs. Let Ω be a domain in $\mathbb{R}^n$ with $n = 1, 2$ or 3. We want to solve the following problem for the states y, z in $Y = Z = H^1(\Omega)$ and the control u in $U = L^2(\Omega)$, where $\hat{y}$ and $\hat{z}$ in $L^2(\Omega)$ are the desired states.

$$\min_{(y,z,u)\in(Y,Z,U)} \frac{1}{2}\int_\Omega (y-\hat{y})^2 + \frac{1}{2}\int_\Omega (z-\hat{z})^2 + \frac{\alpha}{2}\int_\Omega u^2, \tag{2.1}$$

subject to

$$-\epsilon\Delta y + \mathbf{c}\cdot\nabla y = f + u, \qquad \text{in } \Omega, \tag{2.2}$$

$$-\Delta z = g + u, \qquad \text{in } \Omega, \tag{2.3}$$

$$y = \tilde{y}_d, \qquad \text{on } \partial\Omega, \tag{2.4}$$

$$z = \tilde{z}_d, \qquad \text{on } \partial\Omega. \tag{2.5}$$

The data are: f and g are functions in $L^2(\Omega)$ and $\mathbf{c}$ is a vector in $\mathbb{R}^n$.

2.2 Optimize-then-discretize and discretize-then-optimize

This section considers two methods for solving the optimal control problem: *optimize-then-discretize*(OD) and *discretize-then-optimize* (DO). For OD, the optimality conditions for the problem are derived. This yields the state equations, the adjoint equations, as well as an equation that relates the adjoint state variables to the control. Then the collection of equations are solved upon discretization. For DO, first the state PDEs and objective function are discretized. Then the optimality conditions are derived for the discrete optimal control problem. Similarly to the OD method, this yields the discretized state equations, adjoint equations, and an equation that relates the control to the adjoint states.

We use the OD approach to derive the error estimates for our discretization of the problem. For the 1D implementation, we use the DO approach. Note that for the SIPG method for the discretization, both DO and OD are equivalent [11].

2.3 Optimality conditions

2.3.1 Weak form

The weak form of (2.2)-(2.5) is: find y in $H^1(\Omega)$ with $y = \tilde{y_d}$ on $\partial\Omega$ and z in $H^1(\Omega)$ with $z = \tilde{z_d}$ on $\partial\Omega$ such that

$$\int_\Omega (\epsilon \nabla y \cdot \nabla v + (\mathbf{c} \cdot \nabla y)v - fv - uv) = 0, \qquad \forall v \in H_0^1(\Omega),$$
$$\int_\Omega (\nabla z \cdot \nabla q - gq - uq) = 0, \qquad \forall q \in H_0^1(\Omega).$$

Since the variables that will be defined in the Lagrangian need to be in a subspace, the boundary conditions of y and z need to be eliminated. The boundary conditions are lifted by defining y_d in $H^1(\Omega)$ and z_d in $H^1(\Omega)$ such that y_d is equal to $\tilde{y_d}$ on $\partial\Omega$ and z_d is equal to $\tilde{z_d}$ on $\partial\Omega$. We write

$$y = \gamma + y_d,$$
$$z = \eta + z_d.$$

Note that γ and η must vanish on the boundary $\partial\Omega$. The weak formulation can be rewritten as: find γ in $H_0^1(\Omega)$ and η in $H_0^1(\Omega)$ such that

$$\int_\Omega (\epsilon \nabla \gamma \cdot \nabla v + (\mathbf{c} \cdot \nabla \gamma)v - fv - uv) = \int_\Omega (\epsilon \nabla y_d \cdot \nabla v + (\mathbf{c} \cdot \nabla y_d)v), \quad \forall v \in H_0^1(\Omega),$$
$$\int_\Omega (\nabla \eta \cdot \nabla q - gq - uq) = \int_\Omega \nabla z_d \cdot \nabla q, \qquad \forall q \in H_0^1(\Omega).$$

Next, define $\hat{\gamma}$ and $\hat{\eta}$.

$$\hat{\gamma} = \hat{y} - y_d,$$
$$\hat{\eta} = \hat{z} - z_d.$$

Using $\hat{\gamma}$ and $\hat{\eta}$, it follows that the objective function (2.1) can now be written as :

$$\min_{(y,z,u)\in(Y,Z,U)} \frac{1}{2} \int_\Omega \left((y - \hat{y})^2 + (z - \hat{z})^2 + \alpha u^2\right) = \min_{(\gamma,\eta,u)\in(Y_0,Z_0,U)} \frac{1}{2} \int_\Omega \left((\gamma - \hat{\gamma})^2 + (\eta - \hat{\eta})^2 + \alpha u^2\right),$$

where $Y_0 = Z_0 = H_0^1(\Omega)$.

2.3.2 Definition of the Lagrangian

Next, introduce the Lagrangian with the Lagrange multipliers p_y, p_z in $H_0^1(\Omega)$. This is defined by adding the objective function to the weak form of the PDE, where the test functions v and q are replaced with the Lagrange multipliers p_y and p_z, respectively.

$$\begin{aligned} L(\gamma, \eta, u, p_y, p_z) = \frac{1}{2}\int_\Omega \left((\gamma - \hat{\gamma})^2 + (\eta - \hat{\eta})^2 + \alpha u^2\right) \\ + \int_\Omega (\epsilon \nabla\gamma \cdot \nabla p_y + (\mathbf{c} \cdot \nabla\gamma)p_y - f p_y - u p_y) - \int_\Omega (\epsilon \nabla y_d \cdot \nabla p_y + (\mathbf{c} \cdot \nabla y_d)p_y) \\ + \int_\Omega (\nabla\eta \cdot \nabla p_z - g p_z - u p_z) - \int_\Omega \nabla z_d \cdot \nabla p_z. \end{aligned}$$

2.3.3 Derivative of the Lagrangian

The Lagrangian $L(\gamma, \eta, u, p_y, p_z)$ is used to compute the optimality conditions by taking each Frechet derivative of the Lagrangian and setting it equal to zero. These equations will give us the optimality conditions for our problem. Let X be defined as

$$X = H_0^1(\Omega) \times H_0^1(\Omega) \times L^2(\Omega) \times H_0^1(\Omega) \times H_0^1(\Omega).$$

The derivative of L is a linear mapping from X to $\mathbb{R}$. In the rest of this thesis, let the L^2 inner-product on Ω be denoted by $(\cdot, \cdot)_\Omega$. Also, let $\mathbf{n}$ denote the unit normal vector outward to Ω.

Derivative with respect to γ

First, take the Frechet derivative of the Lagrangian with respect to γ and set it equal to zero.

$$\frac{\partial L}{\partial \gamma}(\tilde{\gamma}) = 0, \qquad \forall \tilde{\gamma} \in H_0^1(\Omega).$$

This gives for all $\tilde{\gamma}$ in $H_0^1(\Omega)$:

$$\frac{\partial L}{\partial \gamma}(\tilde{\gamma}) = (\gamma - \hat{\gamma}, \tilde{\gamma})_\Omega + (\epsilon \nabla\tilde{\gamma}, \nabla p_y)_\Omega + (\mathbf{c} \cdot \nabla\tilde{\gamma}, p_y)_\Omega = 0.$$

Using the weak derivatives, we obtain

$$\begin{aligned}\frac{\partial L}{\partial \gamma}(\tilde{\gamma}) &= (\gamma - \hat{\gamma}, \tilde{\gamma})_\Omega + (\epsilon\tilde{\gamma}, \nabla p_y \cdot \mathbf{n})_{\partial\Omega} - (\epsilon\tilde{\gamma}, \Delta p_y)_\Omega \\ &\quad + ((\mathbf{c}\cdot\mathbf{n})\tilde{\gamma}, p_y)_{\partial\Omega} - (\tilde{\gamma}, \mathbf{c}\cdot\nabla p_y)_\Omega \\ &= (\gamma - \hat{\gamma}, \tilde{\gamma})_\Omega - (\epsilon\Delta p_y, \tilde{\gamma})_\Omega - (\mathbf{c}\cdot\nabla p_y, \tilde{\gamma})_\Omega \\ &= (\gamma - \hat{\gamma} - \epsilon\Delta p_y - (\mathbf{c}\cdot\nabla p_y), \tilde{\gamma})_\Omega \\ &= 0.\end{aligned}$$

Since this holds for all $\tilde{\gamma}$ in $H_0^1(\Omega)$,

$$\gamma - \hat{\gamma} - \epsilon\Delta p_y - (\mathbf{c}\cdot\nabla p_y) = 0, \qquad \text{in } \Omega. \tag{2.6}$$

Since $\gamma - \hat{\gamma} = y - \hat{y}$, (2.6) is equivalent to the following:

$$-\epsilon\Delta p_y - \mathbf{c}\cdot\nabla p_y + (y - \hat{y}) = 0, \qquad \text{in } \Omega. \tag{2.7}$$

This optimality condition is the called adjoint equation with respect to state y.

Derivative with respect to η

Next, take the Frechet derivative of the Lagrangian with respect to η and set it equal to zero.

$$\frac{\partial L}{\partial \eta}(\tilde{\eta}) = 0, \qquad \forall\tilde{\eta} \in H_0^1(\Omega).$$

This gives for all $\tilde{\eta}$ in $H_0^1(\Omega)$:

$$\begin{aligned}\frac{\partial L}{\partial \eta}(\tilde{\eta}) &= (\eta - \hat{\eta}, \tilde{\eta})_\Omega + (\nabla\tilde{\eta}, \nabla p_z)_\Omega \\ &= (\eta - \hat{\eta}, \tilde{\eta})_\Omega + (\nabla p_z \cdot \mathbf{n}, \tilde{\eta})_{\partial\Omega} - (\Delta p_z, \tilde{\eta})_\Omega \\ &= (\eta - \hat{\eta}, \tilde{\eta})_\Omega - (\Delta p_z, \tilde{\eta})_\Omega \\ &= (\eta - \hat{\eta} - \Delta p_z, \tilde{\eta})_\Omega \\ &= 0.\end{aligned}$$

Since this holds for all $\tilde{\eta}$ in $H_0^1(\Omega)$, we have

$$\eta - \hat{\eta} - \Delta p_z = 0, \qquad \text{in } \Omega. \tag{2.8}$$

Since $\eta - \hat{\eta} = z - \hat{z}$, (2.8) is equivalent to the following:

$$-\Delta p_z + (z - \hat{z}) = 0, \qquad \text{in } \Omega. \tag{2.9}$$

This optimality condition is the called adjoint equation with respect to state z.

Derivative with respect to u

Now, take the Frechet derivative of the Lagrangian with respect to u and set it equal to zero.

$$\frac{\partial L}{\partial u}(\tilde{u}) = 0, \qquad \forall \tilde{u} \in L^2(\Omega).$$

This gives for all $\tilde{u}$ in $L^2(\Omega)$:

$$\begin{aligned}\frac{\partial L}{\partial u}(\tilde{u}) &= (\alpha u, \tilde{u})_\Omega - (p_y, \tilde{u})_\Omega - (p_z, \tilde{u})_\Omega \\ &= (\alpha u - p_y - p_z, \tilde{u})_\Omega \\ &= 0.\end{aligned}$$

Since this holds for all $\tilde{u}$ in $L^2(\Omega)$, we have

$$\alpha u - p_y - p_z = 0, \qquad \text{in } \Omega. \tag{2.10}$$

Derivative with respect to p_y

Next, take the Frechet derivative of the Lagrangian with respect to p_y and set it equal to zero.

$$\frac{\partial L}{\partial p_y}(\tilde{p}) = 0, \qquad \forall \tilde{p} \in H_0^1(\Omega).$$

This gives for all $\tilde{p}$ in $H_0^1(\Omega)$:

$$\begin{aligned}\frac{\partial L}{\partial p_y}(\tilde{p}) &= (\epsilon\nabla\gamma, \nabla\tilde{p})_\Omega + (\mathbf{c}\cdot\nabla\gamma, \tilde{p})_\Omega - (f, \tilde{p})_\Omega - (u, \tilde{p})_\Omega - (\epsilon\nabla y_d, \nabla\tilde{p})_\Omega - (\mathbf{c}\cdot\nabla y_d, \tilde{p})_\Omega \\ &= (\epsilon\nabla\gamma\cdot\mathbf{n}, \tilde{p})_{\partial\Omega} - (\epsilon\Delta\gamma, \tilde{p})_\Omega - (\epsilon\nabla y_d\cdot\mathbf{n}, \tilde{p})_{\partial\Omega} + (\epsilon\Delta y_d, \tilde{p})_\Omega \\ &+ (\mathbf{c}\cdot\nabla\gamma, \tilde{p})_\Omega - (f, \tilde{p})_\Omega - (u, \tilde{p})_\Omega - (\mathbf{c}\cdot\nabla y_d, \tilde{p})_\Omega \\ &= -(\epsilon\Delta\gamma, \tilde{p})_\Omega + (\epsilon\Delta y_d, \tilde{p})_\Omega + (\mathbf{c}\cdot\nabla\gamma, \tilde{p})_\Omega - (f, \tilde{p})_\Omega - (u, \tilde{p})_\Omega - (\mathbf{c}\cdot\nabla y_d, \tilde{p})_\Omega \\ &= (-\epsilon\Delta\gamma + \mathbf{c}\cdot\nabla\gamma - f - u, \tilde{p})_\Omega + (\epsilon\Delta y_d - \mathbf{c}\cdot\nabla y_d, \tilde{p})_\Omega \\ &= 0.\end{aligned}$$

Since this holds for all $\tilde{p}$ in $H_0^1(\Omega)$, we have

$$-\epsilon\Delta\gamma + \mathbf{c}\cdot\nabla\gamma - f - u = -\epsilon\Delta y_d + \mathbf{c}\cdot\nabla y_d, \qquad \text{in } \Omega.$$

From the definition of γ, we have

$$-\epsilon\Delta y + \mathbf{c}\cdot\nabla y = f + u, \qquad \text{in } \Omega. \tag{2.11}$$

This optimality condition recovers the state equation with respect to y.

Derivative with respect to p_z

Finally, take the Frechet derivative of the Lagrangian with respect to p_z and set it equal to zero.

$$\frac{\partial L}{\partial p_z}(\tilde{p}) = 0, \qquad \forall\tilde{p} \in H_0^1(\Omega).$$

This gives for all $\tilde{p}$ in $H_0^1(\Omega)$:

$$\begin{aligned}\frac{\partial L}{\partial p_z}(\tilde{p}) &= (\nabla\eta, \nabla\tilde{p})_\Omega - (g, \tilde{p})_\Omega - (u, \tilde{p})_\Omega - (\nabla z_d, \nabla\tilde{p})_\Omega \\ &= (\nabla\eta\cdot\mathbf{n}, \tilde{p})_{\partial\Omega} - (\Delta\eta, \tilde{p})_\Omega - (\nabla z_d\cdot\mathbf{n}, \tilde{p})_{\partial\Omega} + (\Delta z_d, \tilde{p})_\Omega - (g, \tilde{p})_\Omega - (u, \tilde{p})_\Omega \\ &= (-\Delta\eta, \tilde{p})_\Omega - (g, \tilde{p})_\Omega - (u, \tilde{p})_\Omega + (\Delta z_d, \tilde{p})_\Omega \\ &= (-\Delta\eta - g - u + \Delta z_d, \tilde{p})_\Omega \\ &= 0.\end{aligned}$$

Since this holds for all $\tilde{p}$ in $H_0^1(\Omega)$, we have

$$-\Delta\eta - g - u = -\Delta z_d, \qquad \text{in } \Omega.$$

From the definition of η, we have

$$-\Delta z = g + u, \qquad \text{in } \Omega. \tag{2.12}$$

This optimality condition recovers the state equation with respect to z.

2.3.4 Optimality conditions

The optimality conditions obtained in the previous section are:

$$\begin{aligned}
-\epsilon\Delta y + \mathbf{c}\cdot\nabla y &= f + u, && \text{in } \Omega,\\
-\Delta z &= g + u, && \text{in } \Omega,\\
-\epsilon\Delta p_y - \mathbf{c}\cdot\nabla p_y &= -(y - \hat{y}), && \text{in } \Omega,\\
-\Delta p_z &= -(z - \hat{z}), && \text{in } \Omega,\\
\alpha u &= p_y + p_z, && \text{in } \Omega,\\
y &= \tilde{y}_d, && \text{on } \partial\Omega,\\
z &= \tilde{z}_d, && \text{on } \partial\Omega,\\
p_y &= 0, && \text{on } \partial\Omega,\\
p_z &= 0, && \text{on } \partial\Omega.
\end{aligned}$$

2.4 Discretization

2.4.1 General notation

Partition the domain $\Omega \subset \mathbb{R}^n$ into N elements denoted by E and denote the mesh E_h, where h is the maximum diameter of the elements. Let Γ_h be the set of interior faces in the mesh. The control and states are approximated using polynomials of

degree less than or equal to k defined on each element E. Let $\mathbb{P}_k(E)$ denote the set of polynomials of degree less than or equal to k on E. Define $V_h(\Omega)$ by

$$V_h(\Omega) = \{v : v \in \mathbb{P}_k(E), \forall E \in E_h\}.$$

Let ϕ_1,...,ϕ_M be a basis of $V_h(\Omega)$, with $M = (k+1)N$. The approximate solutions are denoted y_h, z_h and u_h and can be expanded as follows:

$$y_h(x) = \sum_{i=1}^{(k+1)N} y_i \phi_i(x), \tag{2.13}$$

$$z_h(x) = \sum_{i=1}^{(k+1)N} z_i \phi_i(x), \tag{2.14}$$

$$u_h(x) = \sum_{i=1}^{(k+1)N} u_i \phi_i(x). \tag{2.15}$$

To derive the bilinear form, multiply the PDE by a test function v that lies in $V_h(\Omega)$, integrate by parts over one element, then sum over all the elements. Stabilization terms are added: the penalty term and the symmetrization terms. Interior penalty discontinuous Galerkin methods are used, as well as upwind for the convection term. When using discontinuous Galerkin methods, the jump $[y]$ and the average $\{y\}$ need to be defined on a given face shared by two elements. Let E_1^e and E_2^e be neighboring elements in E_h that have a common face e. Define the normal vector $\mathbf{n}_e$ to be oriented from E_1^e to E_2^e. Then the jump and average are defined as follows:

$$\begin{aligned} [y] &= y|_{E_1^e} - y|_{E_2^e}, \\ \{y\} &= \frac{1}{2}\left(y|_{E_1^e} + y|_{E_2^e}\right). \end{aligned}$$

Define the inflow $\partial\Omega_-$ and the outflow $\partial\Omega_+$ below:

$$\partial\Omega_- = \{x \in \partial\Omega : \mathbf{c} \cdot \mathbf{n}_{\partial\Omega} < 0\},$$

$$\partial\Omega_+ = \{x \in \partial\Omega : \mathbf{c} \cdot \mathbf{n}_{\partial\Omega} \geq 0\}.$$

The binlinear form for the convection term $\mathbf{c} \cdot \nabla y$ is denoted by a_{conv}, and the one for the diffusion terms $-\Delta y$ and $-\Delta z$, is denoted by a_{diff}.

2.4.2 Discretization of the diffusion term

The DG bilinear form for the elliptic operator $-\Delta y$ is:

$$a_{\text{diff}}(y,v) = \sum_{E\in E_h}\int_E \nabla y\cdot\nabla v - \sum_{e\in\Gamma_h\cup\partial\Omega}\int_e \{\nabla y\cdot\mathbf{n}_e\}[v] \tag{2.16}$$

$$+\beta\sum_{e\in\Gamma_h\cup\partial\Omega}\int_e \{\nabla v\cdot\mathbf{n}_e\}[y] + \frac{\sigma_0}{h}\sum_{e\in\Gamma_h\cup\partial\Omega}\int_e [y][v], \quad \forall y,v\in V_h(\Omega). \tag{2.17}$$

The nonsymmetric interior penalty Galerkin Method (NIPG) is obtained, where a_{diff} is nonsymmetric, with $\beta = 1$ and $\sigma_0 = 1$, and symmetric interior penalty Galerkin Method (SIPG) is obtained, where a_{diff} is symmetric, with $\beta = -1$ and σ_0 is bounded below. The bilinear form yields a matrix $\mathbf{A}_{\text{diff}}$ defined as

$$(\mathbf{A}_{\text{diff}})_{ij} = a_{\text{diff}}(\phi_j,\phi_i), \qquad 1\le i,j\le (k+1)N.$$

2.4.3 Discretization of the convection term

Next, the convection term $\mathbf{c}\cdot\nabla y$ is discretized, with the bilinear form a_{conv}:

$$a_{\text{conv}}(y,v) = -\sum_{E\in E_h}\int_E y(\nabla v\cdot\mathbf{c}) + \sum_{e\in\Gamma_h}\int_e y^{\text{up}}[v](\mathbf{c}\cdot\mathbf{n}_e) + \sum_{e\in\partial\Omega_+}\int_e yv(\mathbf{c}\cdot\mathbf{n}_e),$$

where

$$y^{\text{up}} = \begin{cases} y|_{E_1^e}, & \mathbf{c}\cdot\mathbf{n}_e \ge 0, \\ y|_{E_2^e}, & \mathbf{c}\cdot\mathbf{n}_e < 0, \end{cases} \qquad \forall e = \partial E_1^e\cap\partial E_2^e.$$

Lemma 2.1

The following identity holds for all $v\in V_h(\Omega)$ and $w\in H_0^1(\Omega)$:

$$a_{\text{conv}}(v,w) + a_{\text{conv}}(w,v) = 0.$$

Proof 2.1 First we use the definition of a_{conv} to obtain

$$a_{\text{conv}}(v,w)+a_{\text{conv}}(w,v) = -\sum_{E\in E_h}\int_E w(\nabla v\cdot\mathbf{c})+\sum_{e\in\Gamma_h}\int_e w^{\text{up}}[v](\mathbf{c}\cdot\mathbf{n}_e)+\sum_{e\in\partial\Omega_+}\int_e wv(\mathbf{c}\cdot\mathbf{n}_e)$$
$$-\sum_{E\in E_h}\int_E v(\nabla w\cdot\mathbf{c})+\sum_{e\in\Gamma_h}\int_e v^{\text{up}}[w](\mathbf{c}\cdot\mathbf{n}_e)+\sum_{e\in\partial\Omega_+}\int_e vw(\mathbf{c}\cdot\mathbf{n}_e). \tag{2.18}$$

Since w lies in $H_0^1(\Omega)$, $[w]$ is zero almost everywhere and w vanishes on the boundary. We can rewrite (2.18) as

$$\begin{aligned}
&-\sum_{E\in E_h}\int_E w(\nabla v\cdot\mathbf{c})+\sum_{e\in\Gamma_h}\int_e w^{\text{up}}[v](\mathbf{c}\cdot\mathbf{n}_e)+\sum_{e\in\partial\Omega_+}\int_e wv(\mathbf{c}\cdot\mathbf{n}_e)\\
&-\sum_{E\in E_h}\int_E v(\nabla w\cdot\mathbf{c})+\sum_{e\in\Gamma_h}\int_e v^{\text{up}}[w](\mathbf{c}\cdot\mathbf{n}_e)+\sum_{e\in\partial\Omega_+}\int_e vw(\mathbf{c}\cdot\mathbf{n}_e)\\
&\qquad = -\sum_{E\in E_h}\int_E w(\nabla v\cdot\mathbf{c})+\sum_{e\in\Gamma_h}\int_e w^{\text{up}}[v](\mathbf{c}\cdot\mathbf{n}_e)-\sum_{E\in E_h}\int_E v(\nabla w\cdot\mathbf{c}).
\end{aligned} \tag{2.19}$$

Next, we use integration by parts to simplify (2.19).

$$\begin{aligned}
&-\sum_{E\in E_h}\int_E w(\nabla v\cdot\mathbf{c})+\sum_{e\in\Gamma_h}\int_e w^{\text{up}}[v](\mathbf{c}\cdot\mathbf{n}_e)-\sum_{E\in E_h}\int_E v(\nabla w\cdot\mathbf{c})\\
&\qquad = -\sum_{E\in E_h}\int_E w(\nabla v\cdot\mathbf{c})+\sum_{e\in\Gamma_h}\int_e w^{\text{up}}[v](\mathbf{c}\cdot\mathbf{n}_e)\\
&\qquad +\sum_{E\in E_h}\int_E w(\nabla v\cdot\mathbf{c})-\sum_{E\in E_h}\int_{\partial E} vw(\mathbf{c}\cdot\mathbf{n})\\
&\qquad =\sum_{e\in\Gamma_h}\int_e w^{\text{up}}[v](\mathbf{c}\cdot\mathbf{n}_e)-\sum_{E\in E_h}\int_{\partial E} vw(\mathbf{c}\cdot\mathbf{n}_{\partial E}).
\end{aligned} \tag{2.20}$$

We rewrite the integral over ∂E for each edge E in E_h by separating the boundary nodes and interior nodes. (2.20) simplifies to

$$\begin{aligned}
&\sum_{e\in\Gamma_h}\int_e w^{\text{up}}[v](\mathbf{c}\cdot\mathbf{n}_e)-\sum_{E\in E_h}\int_{\partial E} vw(\mathbf{c}\cdot\mathbf{n})\\
&\qquad =\sum_{e\in\Gamma_h}\int_e w^{\text{up}}[v](\mathbf{c}\cdot\mathbf{n}_e)-\sum_{e\in\Gamma_h}\int_e [vw](\mathbf{c}\cdot\mathbf{n}_e)-\sum_{e\in\partial\Omega}\int_e vw(\mathbf{c}\cdot\mathbf{n}_e).
\end{aligned} \tag{2.21}$$

Since $w \in H_0^1(\Omega)$, w vanishes on the boundary, so we can eliminate the boundary term in (2.21).

$$\sum_{e\in\Gamma_h}\int_e w^{\mathrm{up}}[v](\mathbf{c}\cdot\mathbf{n}_e) - \sum_{e\in\Gamma_h}\int_e [vw](\mathbf{c}\cdot\mathbf{n}_e) - \sum_{e\in\partial\Omega}\int_e vw(\mathbf{c}\cdot\mathbf{n}_e)$$
$$= \sum_{e\in\Gamma_h}\int_e (w^{\mathrm{up}}[v] - [vw])(\mathbf{c}\cdot\mathbf{n}_e). \tag{2.22}$$

Since $w \in H_0^1(\Omega)$, we have the following two properties:

$$[vw] = [v]w, \tag{2.23}$$

$$w^{\mathrm{up}} = w. \tag{2.24}$$

Combining (2.23) and (2.24) with (2.22), we complete our proof.

$$\sum_{e\in\Gamma_h}\int_e (w^{\mathrm{up}}[v] - [vw])(\mathbf{c}\cdot\mathbf{n}_e) = \sum_{e\in\Gamma_h}\int_e (w[v] - [v]w)(\mathbf{c}\cdot\mathbf{n}_e) = 0. \tag{2.25}$$

□

This means that for the Lagrange multiplier p_y in $H_0^1(\Omega)$,

$$-a_{\mathrm{conv}}(v, p_y) = a_{\mathrm{conv}}(p_y, v), \quad \forall v \in V_h(\Omega).$$

The bilinear form yields a matrix $\mathbf{A}_{\mathrm{conv}}$ defined as

$$(\mathbf{A}_{\mathrm{conv}})_{ij} = a_{\mathrm{conv}}(\phi_j, \phi_i), \qquad 1 \le i, j \le (k+1)N.$$

2.4.4 Discretization form of the control

Next discretize the control term, u, which appears in both PDEs in the same form.

$$b(u, v) = -\int_\Omega uv.$$

The bilinear form b yields a matrix $\mathbf{B}$ defined as

$$\mathbf{B}_{ij} = b(\phi_j, \phi_i), \qquad 1 \le i, j \le (k+1)N. \tag{2.26}$$

2.4.5 Discretization of the right hand side

Finally, discretize the right hand side for each PDE.

$$l_f(v) = \sum_{E\in E_h}\int_E fv + \sum_{e\in\partial\Omega}\int_e \left(\beta\epsilon\nabla v\cdot\mathbf{n}_e + \frac{\sigma_0}{h}v\right)\tilde{y}_d + \sum_{e\in\partial\Omega_-}\int_e v(\mathbf{c}\cdot\mathbf{n}_e)\tilde{y}_d,$$
$$l_g(v) = \sum_{E\in E_h}\int_E gv + \sum_{e\in\partial\Omega}\int_e \left(\beta\nabla v\cdot\mathbf{n}_e + \frac{\sigma_0}{h}v\right)\tilde{z}_d.$$

These linear forms will yield vectors $\mathbf{f}$ and $\mathbf{g}$, defined by:

$$\mathbf{f}_i = l_f(\phi_i), \quad 1\le i\le (k+1)N,$$
$$\mathbf{g}_i = l_g(\phi_i), \quad 1\le i\le (k+1)N.$$

Assume SIPG is used for all a_{diff} bilinear terms. In particular, note $a_{\text{diff}}(v_h, p_{y_h}) = a_{\text{diff}}(p_{y_h}, v_h)$ and $a_{\text{diff}}(v_h, p_{z_h}) = a_{\text{diff}}(p_{z_h}, v_h)$. The goal is to find $y_h, z_h, u_h, p_{y_h}, p_{z_h}$ in $V_h(\Omega)$ such that for all $v_h \in V_h(\Omega)$

$$\epsilon a_{\text{diff}}(y_h, v_h) + a_{\text{conv}}(y_h, v_h) = (u_h, v_h) + l_f(v_h), \tag{2.27}$$
$$a_{\text{diff}}(z_h, v_h) = (u_h, v_h) + l_g(v_h), \tag{2.28}$$
$$(\alpha u_h, v_h) = (p_{y_h} + p_{z_h}, v_h), \tag{2.29}$$
$$\epsilon a_{\text{diff}}(v_h, p_{y_h}) + a_{\text{conv}}(v_h, p_{y_h}) = -(y_h - \hat{y}, v_h), \tag{2.30}$$
$$a_{\text{diff}}(v_h, p_{z_h}) = -(z_h - \hat{z}, v_h). \tag{2.31}$$

The exact solutions y, z, u, p_y, p_z, also satisfy the system, which is rewritten below.

$$\epsilon a_{\text{diff}}(y, v_h) + a_{\text{conv}}(y, v_h) = (u, v_h) + l_f(v_h), \tag{2.32}$$

$$a_{\text{diff}}(z, v_h) = (u, v_h) + l_g(v_h), \tag{2.33}$$

$$(\alpha u, v_h) = (p_y + p_z, v_h), \tag{2.34}$$

$$\epsilon a_{\text{diff}}(v_h, p_y) + a_{\text{conv}}(v_h, p_y) = -(y - \hat{y}, v_h), \tag{2.35}$$

$$a_{\text{diff}}(v_h, p_z) = -(z - \hat{z}, v_h). \tag{2.36}$$

2.5 *A priori* error estimates

To ensure commutativity of the discretization, which means that *discretize-then-optimize* and *optimize-then-discretize* lead to the same system, SIPG is used for all a_{diff} terms. For more information on the differences of these two approaches, see [11] and [29]. To derive *a priori* error estimates, first define the DG norms.

$$|||v|||^2 = \epsilon |||v|||^2_{\text{diff}} + |||v|||^2_{\text{conv}},$$

$$|||v|||^2_{\text{diff}} = \sum_{E \in E_h} ||\nabla v||^2_{L^2(E)} + \sum_{e \in \Gamma_h \cup \partial\Omega} \frac{1}{h} |||[v]|||^2_{L^2(e)},$$

$$|||v|||^2_{\text{conv}} = \frac{1}{2} \sum_{e \in \Gamma_h} \int_e |\mathbf{c} \cdot \mathbf{n}| [v]^2 + \frac{1}{2} \sum_{e \in \partial\Omega_+} \int_e |\mathbf{c} \cdot \mathbf{n}| v^2.$$

Lemma 2.2

The diffusion discretization a_{diff} as defined in (2.17) is coercive: for all $v \in H^{k+1}(\Omega)$, there exists $C > 0$ such that

$$C|||v|||_{\text{diff}} \leq a_{\text{diff}}(v, v). \tag{2.37}$$

Let C denote a generic constant independent of h that takes different values at different places. Let P_h be the orthogonal L^2 projection defined by

$$(P_h w, v)_E = (w, v)_E, \quad \forall v, w \in V_h(\Omega), \quad \forall E \in E_h.$$

Next, introduce auxiliary functions $\tilde{y}_h$, $\tilde{z}_h$ in $V_h(\Omega)$ that solve the state equations given u,

$$\epsilon a_{\text{diff}}(\tilde{y}_h, v_h) + a_{\text{conv}}(\tilde{y}_h, v_h) = (u, v_h) + l_f(v_h), \quad \forall v_h \in V_h(\Omega), \tag{2.38}$$

$$a_{\text{diff}}(\tilde{z}_h, v_h) = (u, v_h) + l_g(v_h), \quad \forall v_h \in V_h(\Omega). \tag{2.39}$$

Also define $\tilde{p}_{y_h}$, $\tilde{p}_{z_h}$ in $V_h(\Omega)$ that solve the adjoint equations given $\tilde{y}_h$, $\tilde{z}_h$,

$$\epsilon a_{\text{diff}}(v_h, \tilde{p}_{y_h}) + a_{\text{conv}}(v_h, \tilde{p}_{y_h}) = (\hat{y}, v_h) - (\tilde{y}_h, v_h), \quad \forall v_h \in V_h(\Omega), \tag{2.40}$$

$$a_{\text{diff}}(v_h, \tilde{p}_{z_h}) = (\hat{z}, v_h) - (\tilde{z}_h, v_h), \quad \forall v_h \in V_h(\Omega). \tag{2.41}$$

Recall an important result for broken Sobolev spaces. For all w in $H^1(E_h)$, we have

$$||w||_{L^2(\Omega)} \leq C|||w|||_{\text{diff}} \leq \frac{C}{\sqrt{\epsilon}}|||w|||. \tag{2.42}$$

The goal is to bound $||u - u_h||_{L^2(\Omega)}$. Recall from our optimality conditions that $\alpha u = p_y + p_z$.

$$\begin{aligned}
\alpha||u - u_h||^2_{L^2(\Omega)} &= \alpha(u - u_h, u - u_h) \\
&= (\alpha u - \alpha u_h, u - u_h) \pm (p_y - p_{y_h}, u - u_h) \pm (p_z - p_{z_h}, u - u_h) \\
&= (\alpha u - \alpha u_h - (p_y - p_{y_h}) - (p_z - p_{z_h}), u - u_h) \\
&+ (p_y - p_{y_h}, u - u_h) + (p_z - p_{z_h}, u - u_h) \\
&= (\alpha u - p_y - p_z, u - u_h) - (\alpha u_h - p_{y_h} - p_{z_h}, u - u_h) \\
&+ (p_y - p_{y_h}, u - u_h) + (p_z - p_{z_h}, u - u_h).
\end{aligned} \tag{2.43}$$

We use (2.34) and (2.29) to eliminate the term in (2.43), to obtain

$$\begin{aligned}(\alpha u - p_y - p_z, u - u_h) - (\alpha u_h - p_{y_h} - p_{z_h}, u - u_h) \; &+ (p_y - p_{y_h}, u - u_h) + (p_z - p_{z_h}, u - u_h) \\ &= (p_y - p_{y_h}, u - u_h) + (p_z - p_{z_h}, u - u_h) \\ &= (p_y - p_{y_h}, u - u_h) + (p_z - p_{z_h}, u - u_h) \pm (\tilde{p}_{y_h}, u - u_h) \pm (\tilde{p}_{z_h}, u - u_h) \\ &= (p_y - \tilde{p}_{y_h}, u - u_h) + (\tilde{p}_{y_h} - p_{y_h}, u - u_h) \\ &+ (p_z - \tilde{p}_{z_h}, u - u_h) + (\tilde{p}_{z_h} - p_{z_h}, u - u_h) \\ &:= J_{y_1} + J_{y_2} + J_{z_1} + J_{z_2}. \end{aligned} \tag{2.44}$$

Next, bound the terms J_{y_1} and J_{z_1}.

$$\begin{aligned} J_{y_1} &= (p_y - \tilde{p}_{y_h}, u - u_h) \leq \frac{1}{2\frac{\alpha}{2}} ||p_y - \tilde{p}_{y_h}||^2_{L^2(\Omega)} + \frac{\frac{\alpha}{2}}{2} ||u - u_h||^2_{L^2(\Omega)} \\ &\leq \frac{1}{\alpha} ||p_y - \tilde{p}_{y_h}||^2_{L^2(\Omega)} + \frac{\alpha}{4} ||u - u_h||^2_{L^2(\Omega)}, \end{aligned} \tag{2.45}$$

$$\begin{aligned} J_{z_1} &= (p_z - \tilde{p}_{z_h}, u - u_h) \leq \frac{1}{2\frac{\alpha}{2}} ||p_z - \tilde{p}_{z_h}||^2_{L^2(\Omega)} + \frac{\frac{\alpha}{2}}{2} ||u - u_h||^2_{L^2(\Omega)} \\ &\leq \frac{1}{\alpha} ||p_z - \tilde{p}_{z_h}||^2_{L^2(\Omega)} + \frac{\alpha}{4} ||u - u_h||^2_{L^2(\Omega)}. \end{aligned} \tag{2.46}$$

Lemma 2.3

There exists a constant C independent of h and ϵ such that

$$|||\tilde{y}_h - P_h y||| \leq Ch^k |y|_{H^{k+1}(\Omega)}, \tag{2.47}$$

$$|||\tilde{z}_h - P_h z|||_{\text{diff}} \leq Ch^k |z|_{H^{k+1}(\Omega)}. \tag{2.48}$$

This is proven in the Appendix (see Lemma A.4).

Lemma 2.4

There exists a constant C independent of h such that

$$|||\tilde{y}_h - y||| \leq Ch^k |y|_{H^{k+1}(\Omega)} \left(\epsilon^{1/2} + 1 + ||\mathbf{c}||^{1/2}_{\infty} h^{1/2}\right), \tag{2.49}$$

$$|||\tilde{z}_h - z|||_{\text{diff}} \leq Ch^k |z|_{H^{k+1}(\Omega)}. \tag{2.50}$$

Proof 2.2 First we use the triangle inequality to prove (2.49):

$$|||\tilde{y}_h - y||| \leq |||\tilde{y}_h - P_h y||| + |||P_h y - y|||. \tag{2.51}$$

From Lemma A.2, we can bound $|||P_h y - y|||$:

$$|||P_h y - y||| \leq Ch^k |y|_{H^{k+1}(\Omega)} \left(\epsilon^{1/2} + ||\mathbf{c}||_\infty^{1/2} h^{1/2}\right). \tag{2.52}$$

We combine (2.52) and Lemma 2.3 with (2.51), we obtain the desired bound:

$$\begin{aligned} |||\tilde{y}_h - y||| &\leq Ch^k |y|_{H^{k+1}(\Omega)} + Ch^k |y|_{H^{k+1}(\Omega)} \left(\epsilon^{1/2} + ||\mathbf{c}||_\infty^{1/2} h^{1/2}\right) \\ &\leq Ch^k |y|_{H^{k+1}(\Omega)} \left(\epsilon^{1/2} + 1 + ||\mathbf{c}||_\infty^{1/2} h^{1/2}\right). \end{aligned}$$

Next we bound the term (2.50) using the triangle inequality:

$$|||\tilde{z}_h - z|||_{\text{diff}} \leq |||\tilde{z}_h - P_h z|||_{\text{diff}} + |||P_h z - z|||_{\text{diff}}. \tag{2.53}$$

From the error of the L^2 projection, we can bound $|||P_h z - z|||_{\text{diff}}$:

$$|||P_h z - z|||_{\text{diff}} \leq Ch^k |z|_{H^{k+1}(\Omega)}. \tag{2.54}$$

Combining (2.54) and Lemma 2.3 with (2.53), we obtain the desired bound:

$$|||\tilde{z}_h - z|||_{\text{diff}} \leq Ch^k |z|_{H^{k+1}(\Omega)}.$$

□

Lemma 2.5

There exists a constant C independent of h and ϵ such that

$$|||p_y - \tilde{p}_{y_h}|||^2 \leq Ch^{2k} |p_y|^2_{H^{k+1}(\Omega)} \left(\epsilon + ||\mathbf{c}||_\infty h + \epsilon^2 + \frac{||\mathbf{c}||_\infty^2}{\epsilon}\right) + \frac{C}{\epsilon} ||\tilde{y}_h - y||^2_{L^2(\Omega)}, \tag{2.55}$$

$$|||p_z - \tilde{p}_{z_h}|||^2_{\text{diff}} \leq Ch^{2k} |p_z|^2_{H^{k+1}(\Omega)} + C||z - \tilde{z}_h||^2_{L^2(\Omega)}, \tag{2.56}$$

This is proven in the Appendix (see Lemma A.3).

Combining (2.42) and (2.45), we have

$$J_{y_1} \leq \frac{C}{\alpha\epsilon}|||p_y - \tilde{p}_{y_h}|||^2 + \frac{\alpha}{4}||u - u_h||^2_{L^2(\Omega)}. \tag{2.57}$$

Combining (2.57) and Lemma 2.5, we have

$$J_{y_1} \leq Ch^{2k}|p_y|^2_{H^{k+1}(\Omega)}\left(\frac{1}{\alpha} + \frac{||\mathbf{c}||_\infty h}{\alpha\epsilon} + \frac{\epsilon}{\alpha} + \frac{||\mathbf{c}||^2_\infty}{\alpha\epsilon^2}\right) + \frac{C}{\epsilon}||\tilde{y}_h - y||^2_{L^2(\Omega)} + \frac{\alpha}{4}||u - u_h||^2_{L^2(\Omega)}. \tag{2.58}$$

Next, we combine (2.58) with (2.42) and Lemma 2.4 to obtain the bound on J_{y_1}:

$$\begin{aligned}
J_{y_1} &\leq Ch^{2k}|p_y|^2_{H^{k+1}(\Omega)}\left(\frac{1}{\alpha} + \frac{||\mathbf{c}||_\infty h}{\alpha\epsilon} + \frac{\epsilon}{\alpha} + \frac{||\mathbf{c}||^2_\infty}{\alpha\epsilon^2}\right) + \frac{C}{\epsilon^2}|||\tilde{y}_h - y|||^2 + \frac{\alpha}{4}||u - u_h||^2_{L^2(\Omega)} \\
&\leq Ch^{2k}|p_y|^2_{H^{k+1}(\Omega)}\left(\frac{1}{\alpha} + \frac{||\mathbf{c}||_\infty h}{\alpha\epsilon} + \frac{\epsilon}{\alpha} + \frac{||\mathbf{c}||^2_\infty}{\alpha\epsilon^2}\right) + \frac{C}{\epsilon^2}h^{2k}|y|^2_{H^{k+1}(\Omega)}\left(\epsilon + 1 + ||\mathbf{c}||_\infty h\right) \\
&+ \frac{\alpha}{4}||u - u_h||^2_{L^2(\Omega)} \\
&\leq Ch^{2k}|p_y|^2_{H^{k+1}(\Omega)}\left(\frac{1}{\alpha} + \frac{||\mathbf{c}||_\infty h}{\alpha\epsilon} + \frac{\epsilon}{\alpha} + \frac{||\mathbf{c}||^2_\infty}{\alpha\epsilon^2}\right) + Ch^{2k}|y|^2_{H^{k+1}(\Omega)}\left(\frac{1}{\epsilon} + \frac{1 + ||\mathbf{c}||_\infty h}{\epsilon^2}\right) \\
&+ \frac{\alpha}{4}||u - u_h||^2_{L^2(\Omega)}.
\end{aligned} \tag{2.59}$$

Using (2.42) and (2.46), we have

$$J_{z_1} \leq \frac{C}{\alpha}|||p_z - \tilde{p}_{z_h}|||^2_{\text{diff}} + \frac{\alpha}{4}||u - u_h||^2_{L^2(\Omega)}. \tag{2.60}$$

Combining (2.60) and Lemma 2.5, we have

$$J_{z_1} \leq \frac{C}{\alpha}\left(h^{2k}|p_z|^2_{H^{k+1}(\Omega)} + ||z - \tilde{z}_h||^2_{L^2(\Omega)}\right) + \frac{\alpha}{4}||u - u_h||^2_{L^2(\Omega).} \tag{2.61}$$

Next, we combine (2.61) with (2.42) and Lemma 2.4:

$$\begin{aligned}
J_{z_1} &\leq C\left(\frac{h^{2k}}{\alpha}|p_z|^2_{H^{k+1}(\Omega)} + |||z - \tilde{z}_h|||^2_{\text{diff}}\right) + \frac{\alpha}{4}||u - u_h||^2_{L^2(\Omega)} \\
&\leq C\left(\frac{h^{2k}}{\alpha}|p_z|^2_{H^{k+1}(\Omega)} + h^{2k}|z|^2_{H^{k+1}(\Omega)}\right) + \frac{\alpha}{4}||u - u_h||^2_{L^2(\Omega)}.
\end{aligned} \tag{2.62}$$

We have the following equalities for J_{y_2} and J_{z_2}.

Lemma 2.6

$$J_{y_2} = -||\tilde{y}_h - y_h||^2_{L^2(\Omega)}, \tag{2.63}$$

$$J_{z_2} = -||\tilde{z}_h - z_h||^2_{L^2(\Omega)}. \tag{2.64}$$

Proof 2.3 First, we prove the equality for J_{y_2}. We use (2.38) and (2.27) with $v_h = \tilde{p}_{y_h} - p_{y_h}$ to obtain:

$$\begin{aligned} \epsilon a_{\text{diff}}(\tilde{y}_h, \tilde{p}_{y_h} - p_{y_h}) + a_{\text{conv}}(\tilde{y}_h, \tilde{p}_{y_h} - p_{y_h}) &= (u, \tilde{p}_{y_h} - p_{y_h}) + l_f(\tilde{p}_{y_h} - p_{y_h}), & (2.65)\\ \epsilon a_{\text{diff}}(y_h, \tilde{p}_{y_h} - p_{y_h}) + a_{\text{conv}}(y_h, \tilde{p}_{y_h} - p_{y_h}) &= (u_h, \tilde{p}_{y_h} - p_{y_h}) + l_f(\tilde{p}_{y_h} - p_{y_h}) & (2.66) \end{aligned}$$

Subtracting (2.66) from (2.65) gives us the following equation:

$$\epsilon a_{\text{diff}}(\tilde{y}_h - y_h, \tilde{p}_{y_h} - p_{y_h}) + a_{\text{conv}}(\tilde{y}_h - y_h, \tilde{p}_{y_h} - p_{y_h}) = (u - u_h, \tilde{p}_{y_h} - p_{y_h}). \tag{2.67}$$

Next, we use (2.30) and (2.40) with $v_h = \tilde{y}_h - y_h$ to obtain

$$\begin{aligned} \epsilon a_{\text{diff}}(\tilde{y}_h - y_h, \tilde{p}_{y_h}) + a_{\text{conv}}(\tilde{y}_h - y_h, \tilde{p}_{y_h}) &= (\hat{y}, \tilde{y}_h - y_h) - (\tilde{y}_h, \tilde{y}_h - y_h), & (2.68)\\ \epsilon a_{\text{diff}}(\tilde{y}_h - y_h, p_{y_h}) + a_{\text{conv}}(\tilde{y}_h - y_h, p_{y_h}) &= (\hat{y}, \tilde{y}_h - y_h) - (y_h, \tilde{y}_h - y_h). & (2.69) \end{aligned}$$

Subtracting (2.69) from (2.68) gives us the following equation:

$$\begin{aligned} \epsilon a_{\text{diff}}(\tilde{y}_h - y_h, \tilde{p}_{y_h} - p_{y_h}) + a_{\text{conv}}(\tilde{y}_h - y_h, \tilde{p}_{y_h} - p_{y_h}) &= -(\tilde{y}_h - y_h, \tilde{y}_h - y_h)\\ &= -||\tilde{y}_h - y_h||^2_{L^2(\Omega)}. & (2.70) \end{aligned}$$

Combining (2.67) and (2.70), we prove the first part of the lemma.

$$\begin{aligned} J_{y_2} &= (\tilde{p}_{y_h} - p_{y_h}, u - u_h)\\ &= \epsilon a_{\text{diff}}(\tilde{y}_h - y_h, \tilde{p}_{y_h} - p_{y_h}) + a_{\text{conv}}(\tilde{y}_h - y_h, \tilde{p}_{y_h} - p_{y_h})\\ &= -||\tilde{y}_h - y_h||^2_{L^2(\Omega)}. \end{aligned}$$

Next, we prove the equality for J_{z_2}. We use (2.39) and (2.28) with $v_h = \tilde{p}_{z_h} - p_{z_h}$ to get:

$$a_{\text{diff}}(\tilde{z}_h, \tilde{p}_{z_h} - p_{z_h}) = (u, \tilde{p}_{z_h} - p_{z_h}) + l_g(\tilde{p}_{z_h} - p_{z_h}), \tag{2.71}$$

$$a_{\text{diff}}(z_h, \tilde{p}_{z_h} - p_{z_h}) = (u_h, \tilde{p}_{z_h} - p_{z_h}) + l_g(\tilde{p}_{z_h} - p_{z_h}). \tag{2.72}$$

Subtracting (2.72) from (2.71) gives us the following equation:

$$a_{\text{diff}}(\tilde{z}_h - z_h, \tilde{p}_{z_h} - p_{z_h}) = (u - u_h, \tilde{p}_{z_h} - p_{z_h}). \tag{2.73}$$

Next, we use (2.31) and (2.41) with $v_h = \tilde{z}_h - z_h$ to get

$$a_{\text{diff}}(\tilde{z}_h - z_h, \tilde{p}_{z_h}) = (\hat{z}, \tilde{z}_h - z_h) - (\tilde{z}_h, \tilde{z}_h - z_h), \tag{2.74}$$

$$a_{\text{diff}}(\tilde{z}_h - z_h, p_{z_h}) = (\hat{z}, \tilde{z}_h - z_h) - (z_h, \tilde{z}_h - z_h). \tag{2.75}$$

Subtracting (2.75) from (2.74) gives us the following equation:

$$\begin{aligned} a_{\text{diff}}(\tilde{z}_h - z_h, \tilde{p}_{z_h} - p_{z_h}) &= -(\tilde{z}_h - z_h, \tilde{z}_h - z_h) \\ &= -||\tilde{z}_h - z_h||^2_{L^2(\Omega)}. \end{aligned} \tag{2.76}$$

Combining (2.73) and (2.76), we prove the lemma.

$$\begin{aligned} J_{z_2} &= (\tilde{p}_{z_h} - p_{z_h}, u - u_h) \\ &= a_{\text{diff}}(\tilde{z}_h - z_h, \tilde{p}_{z_h} - p_{z_h}) \\ &= -||\tilde{z}_h - z_h||^2_{L^2(\Omega)}. \end{aligned}$$

□

We now return to bounding the term $||u - u_h||_{L^2(\Omega)}$. Using Lemma 2.6 and (2.44), we have

$$\alpha||u - u_h||^2_{L^2(\Omega)} \leq J_{y_1} + J_{z_1}. \tag{2.77}$$

We combine (2.59) and (2.62) with (2.77), we obtain

$$\begin{aligned}
\alpha||u-u_h||^2_{L^2(\Omega)} &\leq Ch^{2k}|p_y|^2_{H^{k+1}(\Omega)}\left(\frac{1}{\alpha}+\frac{||\mathbf{c}||_\infty h}{\alpha\epsilon}+\frac{\epsilon}{\alpha}+\frac{||\mathbf{c}||^2_\infty}{\alpha\epsilon^2}\right)\\
&+ Ch^{2k}|y|^2_{H^{k+1}(\Omega)}\left(\frac{1}{\epsilon}+\frac{1+||\mathbf{c}||_\infty h}{\epsilon^2}\right)+\frac{\alpha}{4}||u-u_h||^2_{L^2(\Omega)}\\
&+ C\left(\frac{h^{2k}}{\alpha}|p_z|^2_{H^{k+1}(\Omega)}+h^{2k}|z|^2_{H^{k+1}(\Omega)}\right)+\frac{\alpha}{4}||u-u_h||^2_{L^2(\Omega)},\\
||u-u_h||^2_{L^2(\Omega)} &\leq Ch^{2k}\left(|y|^2_{H^{k+1}(\Omega)}\left(\frac{1}{\epsilon}+\frac{1+||\mathbf{c}||_\infty h}{\epsilon^2}\right)+|z|^2_{H^{k+1}(\Omega)}\right.\\
&+ \left.|p_y|^2_{H^{k+1}(\Omega)}\left(\frac{1}{\alpha}+\frac{||\mathbf{c}||_\infty h}{\alpha\epsilon}+\frac{\epsilon}{\alpha}+\frac{||\mathbf{c}||^2_\infty}{\alpha\epsilon^2}\right)+\frac{1}{\alpha}|p_z|^2_{H^{k+1}(\Omega)}\right).
\end{aligned}$$

This gives us the error estimates for $||u-u_h||_{L^2(\Omega)}$.

Theorem 2.1

There is a constant C independent of h such that

$$\begin{aligned}
||u-u_h||_{L^2(\Omega)} &\leq Ch^k\left(|y|_{H^{k+1}(\Omega)}\left(\frac{1}{\sqrt{\epsilon}}+\frac{1+||\mathbf{c}||^{1/2}_\infty h^{1/2}}{\epsilon}\right)+|z|_{H^{k+1}(\Omega)}\right.\\
&+ \left.|p_y|_{H^{k+1}(\Omega)}\left(\frac{1}{\sqrt{\alpha}}+\frac{||\mathbf{c}||^{1/2}_\infty h^{1/2}}{\sqrt{\alpha\epsilon}}+\frac{\sqrt{\epsilon}}{\sqrt{\alpha}}+\frac{||\mathbf{c}||_\infty}{\epsilon\sqrt{\alpha}}\right)+\frac{1}{\sqrt{\alpha}}|p_z|_{H^{k+1}(\Omega)}\right).
\end{aligned}$$

Using Theorem 2.1, we can also give error estimates for y and z. We use the triangle inequality to derive the error estimates for z:

$$|||z-z_h|||_{\text{diff}} \leq |||z_h-\tilde{z}_h|||_{\text{diff}}+|||\tilde{z}_h-z|||_{\text{diff}}. \tag{2.78}$$

Using coercivity of a_{diff}, we have

$$C|||z_h-\tilde{z}_h|||^2_{\text{diff}} \leq a_{\text{diff}}(z_h-\tilde{z}_h, z_h-\tilde{z}_h). \tag{2.79}$$

From (2.28) and (2.39), with $v_h = z_h-\tilde{z}_h$, we have

$$a_{\text{diff}}(z_h, z_h-\tilde{z}_h) = (u_h, z_h-\tilde{z}_h)+l_g(z_h-\tilde{z}_h), \tag{2.80}$$

$$a_{\text{diff}}(\tilde{z}_h, z_h-\tilde{z}_h) = (u, z_h-\tilde{z}_h)+l_g(z_h-\tilde{z}_h). \tag{2.81}$$

Subtracting (2.81) from (2.80), we obtain

$$\begin{aligned} a_{\text{diff}}(z_h - \tilde{z}_h, z_h - \tilde{z}_h) &= (u_h - u, z_h - \tilde{z}_h) \\ &\leq ||u - u_h||_{L^2(\Omega)} ||z_h - \tilde{z}_h||_{L^2(\Omega)} \\ &\leq C||u - u_h||_{L^2(\Omega)} |||z_h - \tilde{z}_h|||_{\text{diff}}. \end{aligned} \tag{2.82}$$

Therefore, from (2.79) and (2.82),

$$\begin{aligned} |||z_h - \tilde{z}_h|||^2_{\text{diff}} &\leq C||u - u_h||_{L^2(\Omega)} |||z_h - \tilde{z}_h|||_{\text{diff}}, \\ |||z_h - \tilde{z}_h|||_{\text{diff}} &\leq C||u - u_h||_{L^2(\Omega)}. \end{aligned} \tag{2.83}$$

We can combine (2.83) and (2.78) as well as Theorem 2.1 and Lemma 2.4 to obtain the error estimates for z:

$$|||z - z_h|||_{\text{diff}} \leq Ch^k |z|_{H^{k+1}(\Omega)} + C||u - u_h||_{L^2(\Omega)}. \tag{2.84}$$

Theorem 2.2

There is a constant C independent of h and ϵ such that

$$\begin{aligned} |||z - z_h|||_{\text{diff}} \quad &\leq \quad Ch^k \left(|y|_{H^{k+1}(\Omega)} \left(\frac{1}{\sqrt{\epsilon}} + \frac{1 + ||\mathbf{c}||_\infty^{1/2} h^{1/2}}{\epsilon} \right) + |z|_{H^{k+1}(\Omega)} \right. \\ &+ \quad \left. |p_y|_{H^{k+1}(\Omega)} \left(\frac{1}{\sqrt{\alpha}} + \frac{||\mathbf{c}||_\infty^{1/2} h^{1/2}}{\sqrt{\alpha\epsilon}} + \frac{\sqrt{\epsilon}}{\sqrt{\alpha}} + \frac{||\mathbf{c}||_\infty}{\epsilon\sqrt{\alpha}} \right) + \frac{1}{\sqrt{\alpha}} |p_z|_{H^{k+1}(\Omega)} \right). \end{aligned}$$

Similarly, we derive the error estimates for y. Using the triangle inequality, we have

$$|||y - y_h||| \leq |||y_h - \tilde{y}_h||| + |||\tilde{y}_h - y|||. \tag{2.85}$$

From coercivity of a_{diff}, we have

$$C|||y_h - \tilde{y}_h|||^2 \leq \epsilon a_{\text{diff}}(y_h - \tilde{y}_h, y_h - \tilde{y}_h) + a_{\text{conv}}(y_h - \tilde{y}_h, y_h - \tilde{y}_h). \tag{2.86}$$

From (2.27) and (2.38), with $v_h = y_h - \tilde{y}_h$, we have

$$\epsilon a_{\text{diff}}(y_h, y_h - \tilde{y}_h) + a_{\text{conv}}(y_h, y_h - \tilde{y}_h) = (u_h, y_h - \tilde{y}_h) + l_f(y_h - \tilde{y}_h), \tag{2.87}$$

$$\epsilon a_{\text{diff}}(\tilde{y}_h, y_h - \tilde{y}_h) + a_{\text{conv}}(\tilde{y}_h, y_h - \tilde{y}_h) = (u, y_h - \tilde{y}_h) + l_f(y_h - \tilde{y}_h). \tag{2.88}$$

Subtracting (2.88) from (2.87), we obtain

$$\begin{aligned}\epsilon a_{\text{diff}}(y_h-\tilde{y}_h,y_h-\tilde{y}_h)+a_{\text{conv}}(y_h-\tilde{y}_h,y_h-\tilde{y}_h)&=(u_h-u,y_h-\tilde{y}_h)\\&\leq ||u-u_h||_{L^2(\Omega)}||y_h-\tilde{y}_h||_{L^2(\Omega)}\\&\leq \frac{C}{\sqrt{\epsilon}}||u-u_h||_{L^2(\Omega)}|||y_h-\tilde{y}_h|||.\end{aligned} \tag{2.89}$$

From (2.86) and (2.89), we have

$$\begin{aligned}|||y_h-\tilde{y}_h|||^2&\leq \frac{C}{\sqrt{\epsilon}}||u-u_h||_{L^2(\Omega)}|||y_h-\tilde{y}_h|||,\\|||y_h-\tilde{y}_h|||&\leq \frac{C}{\sqrt{\epsilon}}||u-u_h||_{L^2(\Omega)}.\end{aligned} \tag{2.90}$$

We can combine (2.90) and (2.85) as well as Theorem 2.1 and Lemma 2.4 to obtain the error estimates for y:

$$\begin{aligned}|||y-y_h||| &\leq Ch^k|y|_{H^{k+1}(\Omega)}\left(\epsilon^{1/2}+1+||\mathbf{c}||_\infty^{1/2}h^{1/2}\right)+\frac{C}{\sqrt{\epsilon}}||u-u_h||_{L^2(\Omega)}\\
&\leq Ch^k|y|_{H^{k+1}(\Omega)}\left(\epsilon^{1/2}+1+||\mathbf{c}||_\infty^{1/2}h^{1/2}\right)\\
&+ \frac{C}{\sqrt{\epsilon}}h^k\left(|y|_{H^{k+1}(\Omega)}\left(\frac{1}{\sqrt{\epsilon}}+\frac{1+||\mathbf{c}||_\infty^{1/2}h^{1/2}}{\epsilon}\right)+|z|_{H^{k+1}(\Omega)}\right.\\
&+ \left.|p_y|_{H^{k+1}(\Omega)}\left(\frac{1}{\sqrt{\alpha}}+\frac{||\mathbf{c}||_\infty^{1/2}h^{1/2}}{\sqrt{\alpha\epsilon}}+\frac{\sqrt{\epsilon}}{\sqrt{\alpha}}+\frac{||\mathbf{c}||_\infty}{\epsilon\sqrt{\alpha}}\right)+\frac{1}{\sqrt{\alpha}}|p_z|_{H^{k+1}(\Omega)}\right)\\
&\leq Ch^k\left(|y|_{H^{k+1}(\Omega)}\left(\epsilon^{1/2}+1+||\mathbf{c}||_\infty^{1/2}h^{1/2}+\frac{1}{\sqrt{\epsilon}}\left(\frac{1}{\sqrt{\epsilon}}+\frac{1+||\mathbf{c}||_\infty^{1/2}h^{1/2}}{\epsilon}\right)\right)\right.\\
&+ \frac{1}{\sqrt{\epsilon}}|p_y|_{H^{k+1}(\Omega)}\left(\frac{1}{\sqrt{\alpha}}+\frac{||\mathbf{c}||_\infty^{1/2}h^{1/2}}{\sqrt{\alpha\epsilon}}+\frac{\sqrt{\epsilon}}{\sqrt{\alpha}}+\frac{||\mathbf{c}||_\infty}{\epsilon\sqrt{\alpha}}\right)\\
&+ \left.\frac{1}{\sqrt{\epsilon}}|z|_{H^{k+1}(\Omega)}+\frac{1}{\sqrt{\alpha\epsilon}}|p_z|_{H^{k+1}(\Omega)}\right)\\
&\leq Ch^k\left(|y|_{H^{k+1}(\Omega)}\left(\epsilon^{1/2}+1+||\mathbf{c}||_\infty^{1/2}h^{1/2}+\frac{1}{\epsilon}+\frac{1+||\mathbf{c}||_\infty^{1/2}h^{1/2}}{\epsilon^{3/2}}\right)\right.\\
&+ |p_y|_{H^{k+1}(\Omega)}\left(\frac{1}{\sqrt{\alpha\epsilon}}+\frac{||\mathbf{c}||_\infty^{1/2}h^{1/2}}{\epsilon\sqrt{\alpha}}+\frac{1}{\sqrt{\alpha}}+\frac{||\mathbf{c}||_\infty}{\epsilon\sqrt{\alpha\epsilon}}\right)\\
&+ \left.\frac{1}{\sqrt{\epsilon}}|z|_{H^{k+1}(\Omega)}+\frac{1}{\sqrt{\alpha\epsilon}}|p_z|_{H^{k+1}(\Omega)}\right).\end{aligned}$$

Theorem 2.3

There is a constant C independent of h and ϵ such that

$$\begin{aligned} |||y - y_h||| &\leq Ch^k \Bigg(|y|_{H^{k+1}(\Omega)} \left(\epsilon^{1/2} + 1 + ||\mathbf{c}||_\infty^{1/2} h^{1/2} + \frac{1}{\epsilon} + \frac{1 + ||\mathbf{c}||_\infty^{1/2} h^{1/2}}{\epsilon^{3/2}} \right) \\ &+ |p_y|_{H^{k+1}(\Omega)} \left(\frac{1}{\sqrt{\alpha\epsilon}} + \frac{||\mathbf{c}||_\infty^{1/2} h^{1/2}}{\epsilon\sqrt{\alpha}} + \frac{1}{\sqrt{\alpha}} + \frac{||\mathbf{c}||_\infty}{\epsilon\sqrt{\alpha\epsilon}} \right) \\ &+ \frac{1}{\sqrt{\epsilon}} |z|_{H^{k+1}(\Omega)} + \frac{1}{\sqrt{\alpha\epsilon}} |p_z|_{H^{k+1}(\Omega)} \Bigg). \end{aligned}$$

2.6 Implementation in one dimension

2.6.1 Discretization

For the implementation of this method, *discretize-then-optimize* is used. First, the PDE and the objective function are discretized. Then, the Lagrangian of the discretized problem is used to determine the discrete optimality conditions. Let Ω be the unit interval. Partition the interval with $N + 1$ equidistant nodes, x_n, and let h denote the mesh size.

$$h = \frac{1}{N}, \quad x_n = nh, \quad \forall\, 0 \leq n \leq N.$$

Two piecewise linear functions are used for the basis functions on each interval $[x_{n-1}, x_n]$. The superscripts here denote the basis function number.

$$\forall\, 1 \leq n \leq N, \qquad \phi_n^1(x) = \begin{cases} 1, & x \in [x_{n-1}, x_n], \\ 0, & \text{otherwise}, \end{cases}$$

$$\forall\, 1 \leq n \leq N, \qquad \phi_n^2(x) = \begin{cases} \frac{2}{h} x - \frac{x_{n-1} + x_n}{h}, & x \in [x_{n-1}, x_n], \\ 0, & \text{otherwise}. \end{cases}$$

The following transformation from a general subinterval $[x_{n-1}, x_n]$ to the reference

element, $[-1,1]$ is used.

$$\begin{aligned}
\forall\, 1 \le n \le N, \qquad \hat{x} &= \frac{2}{h}x - \frac{x_{n-1}+x_n}{h}, \\
\forall\, 1 \le n \le N, \qquad \hat{\phi}_n^1(\hat{x}) &= \begin{cases} 1, & \hat{x} \in [-1,1], \\ 0, & \text{otherwise}, \end{cases} \\
\forall\, 1 \le n \le N, \qquad \hat{\phi}_n^2(\hat{x}) &= \begin{cases} \hat{x} & \hat{x} \in [-1,1], \\ 0, & \text{otherwise}. \end{cases}
\end{aligned}$$

The approximate solutions are defined by y_h, z_h and u_h:

$$y_h(x) = \sum_{n=1}^{N}\sum_{i=1}^{2} y_n^i \phi_n^i(x) = \sum_{n=1}^{N}\Big(y_n^1\phi_n^1(x) + y_n^2\phi_n^2(x)\Big), \tag{2.91}$$

$$z_h(x) = \sum_{n=1}^{N}\sum_{i=1}^{2} z_n^i \phi_n^i(x) = \sum_{n=1}^{N}\Big(z_n^1\phi_n^1(x) + z_n^2\phi_n^2(x)\Big), \tag{2.92}$$

$$u_h(x) = \sum_{n=1}^{N}\sum_{i=1}^{2} u_n^i \phi_n^i(x) = \sum_{n=1}^{N}\Big(u_n^1\phi_n^1(x) + u_n^2\phi_n^2(x)\Big). \tag{2.93}$$

Recall the discrete scheme for the state equations:

$$\begin{aligned}
\epsilon a_{\text{diff}}(y_h, v_h) + a_{\text{conv}}(y_h, v_h) &= -b(u_h, v_h) + l_f(v_h), \\
a_{\text{diff}}(z_h, v_h) &= -b(u_h, v_h) + l_g(v_h).
\end{aligned}$$

In one dimension, c is a constant, which is assumed to be positive. The bilinear forms a_{diff}, a_{conv}, and b are

$$
\begin{aligned}
a_{\text{diff}}(y,v) &= \sum_{n=1}^{N}\int_{x_{n-1}}^{x_n} y'v' - \sum_{n=1}^{N-1}\{y'(x_n)\}[v(x_n)] \\
&\quad + \beta\sum_{n=1}^{N-1}\{v'(x_n)\}[y(x_n)] - y'(1)v(1) + y'(0)v(0) + \beta v'(1)y(1) - \beta v'(0)y(0) \\
&\quad + \frac{\sigma_0}{h}\sum_{n=1}^{N-1}[y(x_n)][v(x_n)] + \frac{\sigma_0}{h}y(0)v(0) + \frac{\sigma_0}{h}y(1)v(1), \\
a_{\text{conv}}(y,v) &= -c\sum_{n=1}^{N}\int_{x_{n-1}}^{x_n} yv' + c\sum_{n=1}^{N-1} y(x_n^-)[v(x_n)] + cy(1)v(1), \\
b(u,v) &= -\int_0^1 uv, \\
l_f(v) &= \sum_{n=1}^{N}\int_{x_{n-1}}^{x_n} fv + \frac{\sigma_0}{h}d_{y_0}v(0) + \frac{\sigma_0}{h}d_{y_1}v(1) + \beta\epsilon d_{y_1}v'(1) - \beta\epsilon d_{y_0}v'(0) + cd_{y_0}v(0), \\
l_g(v) &= \sum_{n=1}^{N}\int_{x_{n-1}}^{x_n} gv + \frac{\sigma_0}{h}d_{z_0}v(0) + \frac{\sigma_0}{h}d_{z_1}v(1) + \beta d_{z_1}v'(1) - \beta d_{z_0}v'(0).
\end{aligned}
$$

2.6.2 Implementation

Since piecewise linear polynomials are used, $\mathbf{A}_{\text{diff}}$ lies in $\mathbb{R}^{2N\times 2N}$. It can be shown that $\mathbf{A}_{\text{diff}}$ is defined as follows:

$$
\mathbf{A}_{\text{diff}} = \begin{bmatrix}
\mathbf{F_0} & \mathbf{D} & & & & \\
\mathbf{E} & \mathbf{F} & \mathbf{D} & & & \\
 & \ddots & \ddots & \ddots & & \\
 & & \mathbf{E} & \mathbf{F} & \mathbf{D} & \\
 & & & \mathbf{E} & \mathbf{F_N} &
\end{bmatrix},
$$

where $\mathbf{F}$, $\mathbf{D}$ and $\mathbf{E}$ are defined from the interior nodes, and $\mathbf{F_N}$ and $\mathbf{F_0}$ are defined from the boundary nodes.

$$\mathbf{F} = \frac{1}{h}\begin{bmatrix} 2\sigma_0 & 0 \\ 0 & 2+2\sigma_0+2\beta \end{bmatrix}, \qquad \mathbf{F_0} = \frac{1}{h}\begin{bmatrix} 2\sigma_0 & 1 \\ -\beta & 1+2\sigma_0+3\beta \end{bmatrix},$$

$$\mathbf{F_N} = \frac{1}{h}\begin{bmatrix} 2\sigma_0 & -1 \\ \beta & 1+2\sigma_0+3\beta \end{bmatrix}, \qquad \mathbf{D} = \frac{1}{h}\begin{bmatrix} -\sigma_0 & -1+\sigma_0 \\ -\beta-\sigma_0 & \sigma_0+\beta-1 \end{bmatrix},$$

$$\mathbf{E} = \frac{1}{h}\begin{bmatrix} -\sigma_0 & 1-\sigma_0 \\ \beta+\sigma_0 & \sigma_0+\beta-1 \end{bmatrix}.$$

If $\beta = 1$, the resulting matrix from a_{diff} is non-symmetric, and if $\beta = -1$, $\mathbf{A}_{\text{diff}}$ is symmetric. It can be shown that $\mathbf{A}_{\text{conv}}$, which lies in $\mathbb{R}^{2N\times 2N}$, is defined as follows:

$$\mathbf{A}_{\text{conv}} = \begin{bmatrix} \hat{\mathbf{F}} & & & \\ \hat{\mathbf{E}} & \hat{\mathbf{F}} & & \\ & \ddots & \ddots & \\ & & \hat{\mathbf{E}} & \hat{\mathbf{F}} \end{bmatrix}, \qquad \hat{\mathbf{F}} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \qquad \hat{\mathbf{E}} = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}.$$

It can be shown that $\mathbf{B}$ lies in $\mathbb{R}^{2N\times 2N}$ and is defined as follows:

$$\mathbf{B} = \begin{bmatrix} \hat{\mathbf{B}} & & \\ & \ddots & \\ & & \hat{\mathbf{B}} \end{bmatrix}, \qquad \hat{\mathbf{B}} = -h\begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{bmatrix}.$$

The right hand sides are

$$\mathbf{f} = \mathbf{f_1} + \mathbf{f_2},$$

$$\mathbf{g} = \mathbf{g_1} + \mathbf{g_2},$$

where $\mathbf{f}_1$, $\mathbf{g}_1$ are contributions from f and g, and $\mathbf{f}_2$, $\mathbf{g}_2$ handle the boundary conditions:

$$\mathbf{f_2} = \begin{bmatrix} \frac{\sigma_0 d_{y_0}}{h} + cd_{y_0} \\ -\frac{\sigma_0 d_{y_0}}{h} - \frac{2\beta\epsilon d_{y_0}}{h} - cd_{y_0} \\ 0 \\ \vdots \\ 0 \\ \frac{\sigma_0 d_{y_1}}{h} \\ \frac{\sigma_0 d_{y_1}}{h} + \frac{2\epsilon\beta d_{y_1}}{h} \end{bmatrix}, \qquad \mathbf{g_2} = \begin{bmatrix} \frac{\sigma_0 d_{z_0}}{h} \\ -\frac{\sigma_0 d_{z_0}}{h} - \frac{2\beta d_{z_0}}{h} \\ 0 \\ \vdots \\ 0 \\ \frac{\sigma_0 d_{z_1}}{h} \\ \frac{\sigma_0 d_{z_1}}{h} + \frac{2\beta d_{z_1}}{h} \end{bmatrix}.$$

2.6.3 Discretization of the objective function

Finally, the objective function is discretized. First, discretize one term of the objective function:

$$\begin{aligned} \int_0^1 (y_h(x) - \hat{y}(x))^2 dx &= \sum_{k=1}^{N} \int_{x_{k-1}}^{x_k} (y_h(x) - \hat{y}(x))^2 dx \\ &= \sum_{k=1}^{N} \int_{x_{k-1}}^{x_k} \left(\left(\sum_{n=1}^{N} \sum_{i=1}^{2} y_n^i \phi_n^i(x) \right) - \hat{y}(x) \right)^2 dx. \end{aligned}$$

If i is fixed, ϕ_i^1 and ϕ_i^2, both vanish outside of the interval $[x_{i-1}, x_i]$. Approximate $\hat{y}$ by $\hat{y}_h = P_h \hat{y}$, which can be expanded using the basis functions:

$$\hat{y}_h(x) = \sum_{n=1}^{N} \sum_{i=1}^{2} \hat{y}_n^i \phi_n^i(x).$$

Then this term of the discrete objective function can be defined as

$$\begin{aligned} &\sum_{k=1}^{N} \int_{x_{k-1}}^{x_k} \left(\left(\sum_{n=1}^{N} \sum_{i=1}^{2} y_n^i \phi_n^i(x) \right) - \left(\sum_{n=1}^{N} \sum_{i=1}^{2} \hat{y}_n^i \phi_n^i(x) \right) \right)^2 dx \\ &\qquad = \sum_{k=1}^{N} \int_{x_{k-1}}^{x_k} \left(\sum_{i=1}^{2} \left(y_k^i - \hat{y}_k^i \right) \phi_k^i(x) \right)^2 dx. \end{aligned}$$

Collect the unknowns into vectors $\mathbf{y}$, $\mathbf{z}$, $\mathbf{u}$ that belong to $\mathbb{R}^{2N}$. Recall the superscripts

denote the basis function number and the subscripts denote the interval.

$$\begin{aligned}
\mathbf{y}_{2i-1} &= y_i^1, & \mathbf{y}_{2i} &= y_i^2, & 1 \le i \le N,\\
\mathbf{z}_{2i-1} &= z_i^1, & \mathbf{z}_{2i} &= z_i^2, & 1 \le i \le N,\\
\mathbf{u}_{2i-1} &= u_i^1, & \mathbf{u}_{2i} &= u_i^2, & 1 \le i \le N.
\end{aligned}$$

Let $\mathbf{Q}$ be the mass matrix. From (2.26),

$$\mathbf{Q} = -\mathbf{B}.$$

The first term of the objective function is rewritten as

$$\frac{1}{2}\int_0^1 (y_h - \hat{y}_h)^2 = \frac{1}{2}\left(\mathbf{y} - \hat{\mathbf{y}}\right)^T \mathbf{Q}\left(\mathbf{y} - \hat{\mathbf{y}}\right).$$

This process can be applied to the rest of the objective function. Define $\mathbf{R}$ to be equal to $\mathbf{Q}$. We use a different variable because the control u_h is not necessarily in the same space as the states y_h and z_h. In this problem, the spaces are the same, so the discretization matrices $\mathbf{R}$ and $\mathbf{Q}$ are the same.

$$\frac{1}{2}\int_0^1 (z_h - \hat{z}_h)^2 = \frac{1}{2}\left(\mathbf{z} - \hat{\mathbf{z}}\right)^T \mathbf{Q}\left(\mathbf{z} - \hat{\mathbf{z}}\right),$$

$$\frac{\alpha}{2}\int_0^1 u_h^2 = \frac{\alpha}{2}\mathbf{u}^T\mathbf{R}\mathbf{u}.$$

2.6.4 Fully discretized form

The fully discretized problem has become

$$\min_{(\mathbf{y},\mathbf{z},\mathbf{u})\in(\mathbb{R}^{2N},\mathbb{R}^{2N},\mathbb{R}^{2N})} \left(\frac{1}{2}(\mathbf{z}-\hat{\mathbf{z}})^T\mathbf{Q}(\mathbf{z}-\hat{\mathbf{z}}) + \frac{1}{2}(\mathbf{y}-\hat{\mathbf{y}})^T\mathbf{Q}(\mathbf{y}-\hat{\mathbf{y}}) + \frac{\alpha}{2}\mathbf{u}^T\mathbf{R}\mathbf{u},\right)$$

subject to

$$\begin{aligned}
\epsilon\mathbf{A}_{\text{diff}}\mathbf{y} + \mathbf{A}_{\text{conv}}\mathbf{y} &= -\mathbf{B}\mathbf{u} + \mathbf{f},\\
\mathbf{A}_{\text{diff}}\mathbf{z} &= -\mathbf{B}\mathbf{u} + \mathbf{g}.
\end{aligned}$$

2.7 Optimization using the discrete Lagrangian

To solve the optimal control problem, the process outlined in [17] is followed. Define the discrete Lagrangian and the Lagrange multipliers $\mathbf{p}_y$ in $\mathbb{R}^{2N}$ and $\mathbf{p}_z$ in $\mathbb{R}^{2N}$, which correspond to the states $\mathbf{y}$ and $\mathbf{z}$ respectively.

$$\begin{aligned} L(\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{p}_y,\mathbf{p}_z) &= \frac{1}{2}(\mathbf{z}-\hat{\mathbf{z}})^T\mathbf{Q}(\mathbf{z}-\hat{\mathbf{z}}) + \frac{1}{2}(\mathbf{y}-\hat{\mathbf{y}})^T\mathbf{Q}(\mathbf{y}-\hat{\mathbf{y}}) \\ &+ \frac{\alpha}{2}\mathbf{u}^T\mathbf{R}\mathbf{u} + \mathbf{p}_y^T(\epsilon\mathbf{A}_{\text{diff}}\mathbf{y} + \mathbf{A}_{\text{conv}}\mathbf{y} + \mathbf{B}\mathbf{u} - \mathbf{f}) \\ &+ \mathbf{p}_z^T(\mathbf{A}_{\text{diff}}\mathbf{z} + \mathbf{B}\mathbf{u} - \mathbf{g}). \end{aligned}$$

To determine the optimality conditions, we take the gradient of the Lagrangian and set it equal to zero.

$$\begin{aligned} \nabla_{\mathbf{y}} L &= \mathbf{Q}(\mathbf{y}-\hat{\mathbf{y}}) + (\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{diff}})^T\mathbf{p}_y &&= \mathbf{0}, \\ \nabla_{\mathbf{z}} L &= \mathbf{Q}(\mathbf{z}-\hat{\mathbf{z}}) + \mathbf{A}_{\text{diff}}^T\mathbf{p}_z &&= \mathbf{0}, \\ \nabla_{\mathbf{u}} L &= \alpha\mathbf{R}\mathbf{u} + \mathbf{B}^T\mathbf{p}_y + \mathbf{B}^T\mathbf{p}_z &&= \mathbf{0}, \\ \nabla_{\mathbf{p}_y} L &= \epsilon\mathbf{A}_{\text{diff}}\mathbf{y} + \mathbf{A}_{\text{conv}}\mathbf{y} + \mathbf{B}\mathbf{u} - \mathbf{f} &&= \mathbf{0}, \\ \nabla_{\mathbf{p}_z} L &= \mathbf{A}_{\text{diff}}\mathbf{z} + \mathbf{B}\mathbf{u} - \mathbf{g} &&= \mathbf{0}. \end{aligned}$$

This gives us a linear system of five variables and five unknowns, thus we can solve the following system with a linear solver.

$$\begin{bmatrix} \mathbf{Q} & \mathbf{0} & \mathbf{0} & (\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}})^T & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{\text{diff}}^T \\ \mathbf{0} & \mathbf{0} & \alpha\mathbf{R} & \mathbf{B}^T & \mathbf{B}^T \\ \epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}} & \mathbf{0} & \mathbf{B} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{\text{diff}} & \mathbf{B} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \\ \mathbf{u} \\ \mathbf{p}_y \\ \mathbf{p}_z \end{bmatrix} = \begin{bmatrix} \mathbf{Q}\hat{\mathbf{y}} \\ \mathbf{Q}\hat{\mathbf{z}} \\ \mathbf{0} \\ \mathbf{f} \\ \mathbf{g} \end{bmatrix}.$$

2.8 Numerical examples

We solve the linear system using code written in Matlab. We use the exact solutions to create the synthetic data to make up $\hat{y}$, $\hat{z}$ and the right hand sides f and g. For

all the examples, the following constants are used:

$$\alpha = 1, \quad \epsilon = 10^{-9}, \quad c = 1.$$

Recall that α is the coefficient of u in the objective function, ϵ is the diffusion coefficient, and c is the convection coefficient in the y equation. Recall from the optimality conditions the following property of u:

$$u(x) = \frac{1}{\alpha}(p_y(x) + p_z(x)), \quad \forall\, x \in (0,1).$$

For all the examples, we increase the number of intervals the same way:

$$N = 32, 64, 128.$$

The figures report the errors validating the estimates proved in Section 2.5. $L^2(0,1)$ errors are also reported. Recall that σ_0 is the penalty parameter defined when using NIPG or SIPG. In the following examples, when using SIPG, let $\sigma_0 = 10$ and when using NIPG, let $\sigma_0 = 1$. Define the L^2 error for y_h as $||y - y_h||_{L^2(\Omega)}$ and the energy error for y_h as $\left(\sum_{E \in E_h} ||\nabla(y - y_h)||^2_{L^2(E)}\right)^{1/2}$.

Example 1 uses linear basis functions, DG SIPG, and the exact solutions below:

$$y(x) = x^2, \qquad z(x) = \cos(x),$$

$$p_y(x) = x^3 - x^4, \qquad p_z(x) = x^4 - x^5, \qquad u(x) = x^3 - x^5.$$

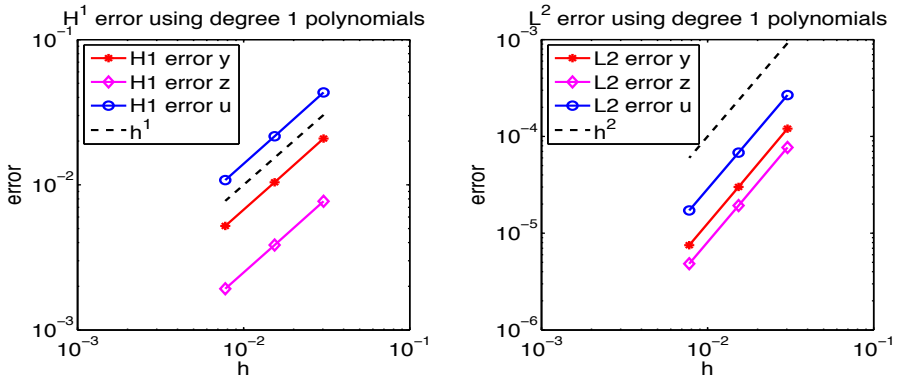


Figure 2.1 : H^1 versus h (left), L^2 versus h (right) for example 1.

Example 2 uses linear basis functions, DG SIPG, and the exact solutions below:

$$y(x) = e^{-x^2} + 1, \qquad z(x) = x^2\cos(x) + x,$$

$$p_y(x) = x^2(1-x)\sin(x), \qquad p_z(x) = x^4(1-x), \qquad u(x) = x^2(1-x)\left(\sin(x) + x^2\right).$$

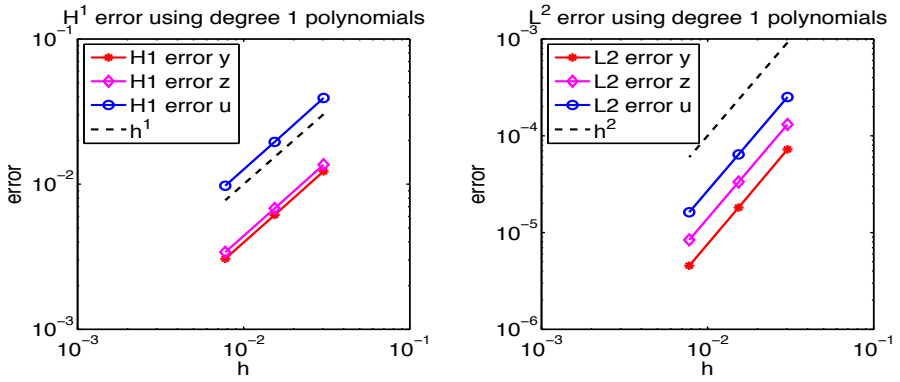


Figure 2.2 : H^1 versus h (left), L^2 versus h (right) for example 2.

Example 3 solves the problem using quadratic basis functions with DG SIPG, with exact solutions from example 2 (Figure 2.2).

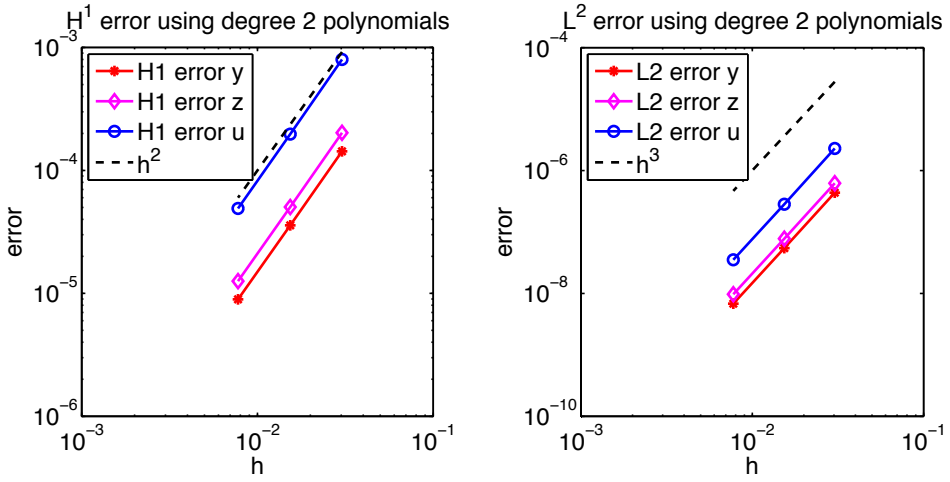


Figure 2.3 : H^1 versus h (left), L^2 versus h (right) for example 5.

Example 4 solves the problem using quadratic basis functions with DG NIPG, with exact solutions from example 2 (Figure 2.2).

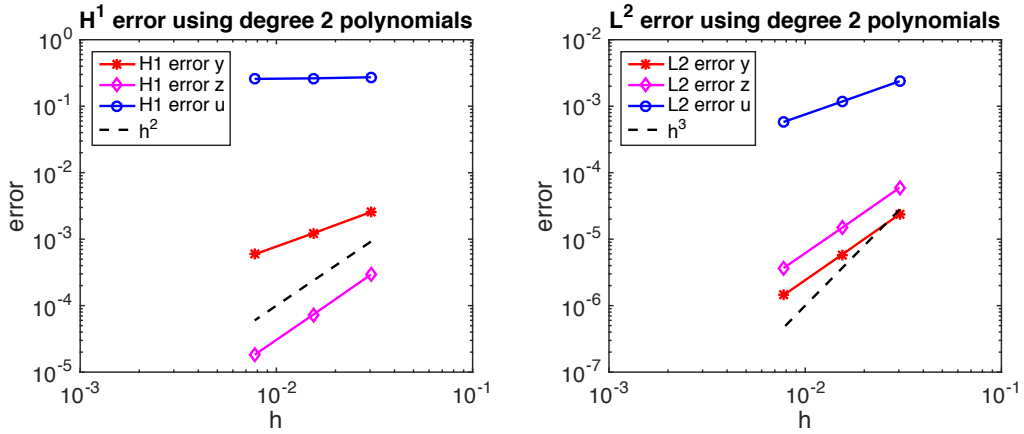


Figure 2.4 : H^1 versus h (left), L^2 versus h (right) for example 4.

Example 5 solves the problem using linear basis functions with DG SIPG, with the following exact solutions:

$$y(x) = x^3 - \frac{e^{(x-1)/\epsilon} - e^{-1/\epsilon}}{1 - e^{-1/\epsilon}}, \qquad z(x) = x^2\cos(x) + x,$$

$$p_y(x) = x^2(1-x)\sin(x), \qquad p_z(x) = x^4(1-x), \qquad u(x) = x^2(1-x)\left(\sin(x) + x^2\right).$$

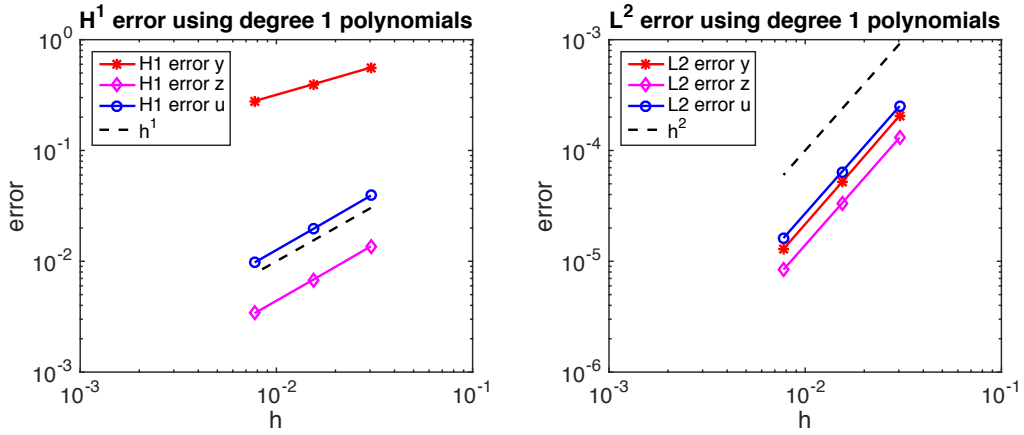


Figure 2.5 : H^1 versus h (left), L^2 versus h (right) for example 2.

From the examples, it can be observed that DG SIPG works well. Using linear basis functions, both states y, z and the control u all have convergence rates of two in the L^2 norm and one in the energy norm, which is observed in Figures 2.1-2.2. When using quadratic basis polynomials and SIPG, the L^2 error is order h^3 and energy error is order h^2 for y, z, u, which is observed in Figure 2.3. These optimal rates are achieved because SIPG is commutative.

In Figure 2.5, the exact solution contains boundary layers, which can cause problems in convergence. It can be observed that the convergence rate of y in the H^1 norm is .5, which is not the optimal rate of one. It can be noted that y converges optimally in the L^2 norm at the rate of two. The presence of boundary layers in y affects its convergence rates, but not the convergence rates of the other variables, as

the rates in both norms, for z and u are optimal, i.e. one in the H^1 norm and two in the L^2 norm.

When using NIPG, the two different approaches do not lead to the same system, and as observed in Figure 2.4, the rates are not optimal. Though y and z converge in both norms, u does not converge in the energy norm. The error estimates proven in this chapter only apply to SIPG, so we do not necessarily have the same error estimates when using NIPG. Our examples illustrate that DG works well for convection-dominated PDEs. A small diffusion coefficient was chosen to show the convergence rates are still optimal when the diffusion term is small.

Chapter 3

Optimal Control of the Transport Equation

In this chapter, we solve an optimal control problem governed by the transport equation. First, the continuous optimality conditions are derived. Next, the PDE and objective function are discretized using DG methods in space and the trapezoid method in time. The discrete optimality conditions are derived using the discrete Lagrangian. To solve the minimization problem, we use the Newton-Conjugate Gradient method is used. Last, numerical examples are provided to validate the accuracy of the methods.

3.1 Problem statement

The section focuses on solving an optimal control problem governed by a convection-dominated transport equation. Define the state space Y and the control space U to be:

$$Y = \{y \in L^2(0,T;H^1(\Omega)), y_t \in L^2(0,T;H^1(\Omega))\},$$
$$U = L^2(0,T;L^2(\Omega)).$$

The goal is to solve the following problem for the state y in Y and the control u in U. In the following problem, $\hat{y}$ which lies in $L^2(0,T;L^2(\Omega))$ is the desired state.

$$\min_{(y,u)\in(Y,U)} \frac{1}{2}\int_0^T \int_\Omega (y-\hat{y})^2 + \frac{\alpha}{2}\int_0^T \int_\Omega u^2, \tag{3.1}$$

subject to

$$
\begin{aligned}
y_t - \epsilon\Delta y + \mathbf{c}\cdot\nabla y &= f + u, \qquad \text{in } \Omega \times (0,T], && (3.2)\\
y &= \tilde{y}_d, \qquad \text{on } \partial\Omega \times (0,T], && (3.3)\\
y(t=0) &= y_0, \qquad \text{in } \Omega. && (3.4)
\end{aligned}
$$

3.2 Optimality conditions

3.2.1 Weak form

The weak form of (3.2) - (3.4) is: Find y in Y such that

$$\int_0^T \int_\Omega (y_t v + \epsilon\nabla y\cdot\nabla v + (\mathbf{c}\cdot\nabla y)v - fv - uv) = 0, \qquad \forall v \in L^2(0,T;H^1(\Omega)).$$

The boundary conditions are lifted by defining γ in $L^2(0,T;H^1(\Omega))$ with γ_t in $L^2(0,T;H^1(\Omega))$ such that γ is equal to $\tilde{y}_d$ on $\partial\Omega$. We write

$$y = \gamma + y_d.$$

Then we can see that γ vanishes on the boundary $\partial\Omega$. Let Y_0 be defined by:

$$Y_0 = \{\gamma \in L^2(0,T;H_0^1(\Omega)), \gamma_t \in L^2(0,T;H_0^1(\Omega))\},$$

Now the weak form can be rewritten as: find γ in Y_0 such that

$$
\begin{aligned}
&\int_0^T \int_\Omega (\gamma_t v + \epsilon\nabla\gamma\cdot\nabla v + (\mathbf{c}\cdot\nabla\gamma)v - fv - uv)\\
&\qquad = \int_0^T \int_\Omega \left((y_d)_t\, v + \epsilon\nabla y_d\cdot\nabla v + (\mathbf{c}\cdot\nabla y_d)v\right), \quad \forall v \in L^2(0,T;H_0^1(\Omega)).
\end{aligned}
$$

The initial condition for γ using y_0 in Ω is given by

$$
\begin{aligned}
\gamma_0(\cdot) &:= \gamma(\cdot,0)\\
&= y(\cdot,0) - y_d(\cdot,0)\\
&= y_0(\cdot) - y_d(\cdot,0).
\end{aligned}
$$

Define $\hat{\gamma}$:

$$\hat{\gamma} = \hat{y} - y_d.$$

It follows that the objective function can now be written as:

$$\min_{(y,u)\in(Y,U)} \frac{1}{2}\int_0^T \int_\Omega \left((y-\hat{y})^2 + \alpha u^2\right) = \min_{(\gamma,u)\in(Y_0,U)} \frac{1}{2}\int_0^T \int_\Omega \left((\gamma-\hat{\gamma})^2 + \alpha u^2\right).$$

We also weakly enforce initial condition with the following:

$$\int_\Omega \left(y(\cdot,0) - y_0\right) v_0 = 0, \qquad \forall v_0 \in L^2(\Omega).$$

3.2.2 Definition of the Lagrangian

Next, introduce the Lagrangian with the Lagrange multiplier p, which lies in Y_0. This is defined by adding the objective function to the weak form of the PDE, where the test function v is replaced by the Lagrange multiplier p.

$$\begin{aligned} L(\gamma,u,p) = &\frac{1}{2}\int_0^T \int_\Omega \left((\gamma-\hat{\gamma})^2 + \alpha u^2\right) \\ &+ \int_0^T \int_\Omega \left(\gamma_t p + \epsilon \nabla\gamma\cdot\nabla p + (\mathbf{c}\cdot\nabla\gamma)p - fp - up\right) \\ &- \int_0^T \int_\Omega \left((y_d)_t p + \epsilon\nabla y_d\cdot\nabla p + (\mathbf{c}\cdot\nabla y_d)p\right) \\ &+ \int_\Omega \left(\gamma(\cdot,0) - \gamma_0\right) p_0. \end{aligned}$$

3.2.3 Derivative of the Lagrangian

The Lagrangian $L(\gamma,u,p)$ is used to compute the optimality conditions. Taking a derivative of the Lagrangian with respect to each variable and setting it equal to zero determines the optimality conditions.

Derivative with respect to γ

Take the derivative of the Lagrangian with respect to γ and set it equal to zero.

$$\frac{\partial L}{\partial \gamma}(\tilde{\gamma}) = 0, \qquad \forall \tilde{\gamma} \in Y_0.$$

We have for all $\tilde{\gamma}$ in Y_0,

$$\begin{aligned}
\frac{\partial L}{\partial \gamma}(\tilde{\gamma}) &= \int_0^T ((\gamma - \hat{\gamma}, \tilde{\gamma})_\Omega + (\tilde{\gamma}_t, p)_\Omega + \epsilon(\nabla\tilde{\gamma}, \nabla p)_\Omega + (\mathbf{c}\cdot\nabla\tilde{\gamma}, p)_\Omega) + (\tilde{\gamma}(\cdot,0), p_0)_\Omega \\
&= \int_0^T ((\gamma - \hat{\gamma}, \tilde{\gamma})_\Omega - (p_t, \tilde{\gamma})_\Omega + \epsilon(\nabla p\cdot\mathbf{n}, \tilde{\gamma})_{\partial\Omega} - (\epsilon\Delta p, \tilde{\gamma})_\Omega + ((\mathbf{c}\cdot\mathbf{n})p, \tilde{\gamma})_{\partial\Omega} - (\mathbf{c}\cdot\nabla p, \tilde{\gamma})_\Omega) \\
&+ \int_\Omega (p(\cdot,T)\tilde{\gamma}(\cdot,T) - p(\cdot,0)\tilde{\gamma}(\cdot,0)) + (\tilde{\gamma}(\cdot,0), p_0)_\Omega \\
&= \int_0^T ((\gamma - \hat{\gamma}, \tilde{\gamma})_\Omega - (p_t, \tilde{\gamma})_\Omega - (\epsilon\Delta p, \tilde{\gamma})_\Omega - (\mathbf{c}\cdot\nabla p, \tilde{\gamma})_\Omega) + \int_\Omega (p(\cdot,T)\tilde{\gamma}(\cdot,T)) \\
&= 0.
\end{aligned}$$

We now choose p such that $p(\cdot,T)$ vanishes in Ω. Then we have for all p in Y_0

$$\int_0^T ((\gamma - \hat{\gamma}, \tilde{\gamma})_\Omega - (p_t, \tilde{\gamma})_\Omega - (\epsilon\Delta p, \tilde{\gamma})_\Omega - (\mathbf{c}\cdot\nabla p, \tilde{\gamma})_\Omega) = 0.$$

Since this holds for all p in Y_0, we have

$$(\gamma - \hat{\gamma}) - p_t - \epsilon\Delta p - \mathbf{c}\cdot\nabla p = 0, \qquad \text{in } \Omega \times (0,T).$$

Since $\gamma - \hat{\gamma} = y - \hat{y}$, the above PDE is equivalent to the following:

$$\begin{aligned}
-p_t - \epsilon\Delta p - \mathbf{c}\cdot\nabla p + (y - \hat{y}) = 0, &\qquad \text{in } \Omega \times (0,T), \\
p(\cdot,T) = 0, &\qquad \text{in } \Omega, \\
-p(\cdot,0) + p_0 = 0, &\qquad \text{in } \Omega.
\end{aligned}$$

Derivative with respect to u

Next, take the derivative of the Lagrangian with respect to u and set it equal to zero.

$$\frac{\partial L}{\partial u}(\tilde{u}) = 0, \qquad \forall \tilde{u} \in U.$$

We have for all $\tilde{u}$ in U,

$$\frac{\partial L}{\partial u}(\tilde{u}) = \int_0^T (\alpha(u, \tilde{u})_\Omega - (p, \tilde{u})_\Omega) = 0.$$

Since this holds for all $\tilde{u}$, this implies

$$\alpha u - p = 0, \qquad \text{in } \Omega \times (0, T).$$

Derivative with respect to p

Then, take the derivative of the Lagrangian with respect to p and set it equal to zero.

$$\frac{\partial L}{\partial p}(\tilde{p}) = 0, \qquad \forall \tilde{p} \in Y_0.$$

We have for all $\tilde{p}$ in Y_0,

$$\begin{aligned}
\frac{\partial L}{\partial p}(\tilde{p}) &= \int_0^T ((\gamma_t, \tilde{p})_\Omega + \epsilon(\nabla\gamma, \nabla\tilde{p})_\Omega + (\mathbf{c}\cdot\nabla\gamma, \tilde{p})_\Omega - (f, \tilde{p})_\Omega - (u, \tilde{p})_\Omega) \\
&\quad - \int_0^T (((y_d)_t, \tilde{p})_\Omega + \epsilon(\nabla y_d, \nabla\tilde{p})_\Omega + (\mathbf{c}\nabla y_d, \tilde{p})_\Omega) \\
&= \int_0^T ((\gamma_t, \tilde{p})_\Omega + \epsilon(\nabla\gamma\cdot\mathbf{n}, \tilde{p})_{\partial\Omega} - \epsilon(\Delta\gamma, \tilde{p})_\Omega + (\mathbf{c}\cdot\nabla\gamma, \tilde{p})_\Omega - (f, \tilde{p})_\Omega - (u, \tilde{p})_\Omega) \\
&\quad - \int_0^T (((y_d)_t, \tilde{p})_\Omega + \epsilon(\nabla y_d\cdot\mathbf{n}, \tilde{p})_{\partial\Omega} - \epsilon(\Delta y_d, \tilde{p})_\Omega + (\mathbf{c}\nabla y_d, \tilde{p})_\Omega) \\
&= \int_0^T ((\gamma_t, \tilde{p})_\Omega - \epsilon(\Delta\gamma, \tilde{p})_\Omega + (\mathbf{c}\cdot\nabla\gamma, \tilde{p})_\Omega - (f, \tilde{p})_\Omega - (u, \tilde{p})_\Omega) \\
&\quad - \int_0^T (((y_d)_t, \tilde{p})_\Omega - \epsilon(\Delta y_d, \tilde{p})_\Omega + (\mathbf{c}\nabla y_d, \tilde{p})_\Omega) \\
&= 0.
\end{aligned}$$

Since this holds for all $\tilde{p}$ in Y, this implies

$$\gamma_t - \epsilon\Delta\gamma + \mathbf{c}\cdot\nabla\gamma - f - u = (y_d)_t - \epsilon\Delta y_d + \mathbf{c}\cdot\nabla y_d, \qquad \text{in } \Omega \times (0, T).$$

By the definition of γ, we now have

$$y_t - \epsilon\Delta y + \mathbf{c}\cdot\nabla y = f + u, \qquad \text{in } \Omega \times (0, T).$$

Derivative with respect to p_0

Last, take the derivative of the Lagrangian with respect to p_0 and set it equal to zero.

$$\frac{\partial L}{\partial p_0}(\tilde{p}_0) = 0, \qquad \forall \tilde{p} \in L^2(\Omega).$$

We have for all $\tilde{p}_0$ in $L^2(\Omega)$,

$$\frac{\partial L}{\partial p_0}(\tilde{p}_0) = (\gamma(\cdot, 0) - \gamma_0, \tilde{p}_0)_\Omega = 0.$$

Since this holds for all p_0 in $L^2(\Omega)$,

$$\gamma(\cdot, 0) - \gamma_0 = 0, \quad \text{in } \Omega.$$

This is equivalent to the following

$$y(\cdot, 0) - y_0 = 0, \quad \text{in } \Omega.$$

3.2.4 Optimality conditions

The optimality conditions from the previous section are summarized below.

$$\begin{aligned}
y_t - \epsilon \Delta y + \mathbf{c} \cdot \nabla y &= f + u, && \text{in } \Omega \times (0, T), \\
-p_t - \epsilon \Delta p - \mathbf{c} \cdot \nabla p &= -(y - \hat{y}) && \text{in } \Omega \times (0, T), \\
\alpha u &= p, && \text{in } \Omega \times (0, T), \\
y &= y_d, && \text{on } \partial\Omega \times (0, T), \\
p &= 0, && \text{on } \partial\Omega \times (0, T), \\
y(\cdot, 0) &= y_0, && \text{in } \Omega, \\
p(\cdot, 0) &= p_0, && \text{in } \Omega, \\
p(\cdot, T) &= 0, && \text{in } \Omega.
\end{aligned}$$

3.3 Discretization

To solve the optimal control problem (3.1)-(3.4), we use the *discretize-then-optimize* approach [29]. The same discretization in space as in Section 2.4 is utilized. Recall this discretization defined by:

$$
\begin{aligned}
a_{\text{diff}}(y,v) &= \sum_{E\in E_h}\int_E \nabla y\cdot\nabla v - \sum_{e\in\Gamma_h\cup\partial\Omega}\int_e \{\nabla y\cdot\mathbf{n}_e\}[v] \\
&\quad + \beta\sum_{e\in\Gamma_h\cup\partial\Omega}\int_e \{\nabla v\cdot\mathbf{n}_e\}[y] + \frac{\sigma_0}{h}\sum_{e\in\Gamma_h\cup\partial\Omega}\int_e [y][v], \quad \forall y,v\in V_h(\Omega), \\
a_{\text{conv}}(y,v) &= -\sum_{E\in E_h}\int_E y(\nabla v\cdot\mathbf{c}) + \sum_{e\in\Gamma_h}\int_e y^{\text{up}}(\mathbf{c}\cdot\mathbf{n}_e)[v] + \sum_{e\in\partial\Omega_+}\int_e yv(\mathbf{c}\cdot\mathbf{n}_{\partial\Omega}), \\
b(u,v) &= -\int_\Omega uv, \\
l_f(v) &= \sum_{E\in E_h}\int_E fv + \sum_{e\in\partial\Omega}\int_e \left(\beta\epsilon\nabla v\cdot\mathbf{n}_e + \frac{\sigma_0}{h}v\right)\tilde{y}_d + \sum_{e\in\partial\Omega_-}\int_e v(\mathbf{c}\cdot\mathbf{n}_e)\tilde{y}_d.
\end{aligned}
$$

The goal is to find v in $V_h(\Omega)$ such that

$$
(y_t,v)_\Omega + \epsilon a_{\text{diff}}(y,v) + a_{\text{conv}}(y,v) = b(u,v) + l_f(v). \tag{3.5}
$$

The time derivative term, $(y_t,v)_\Omega$, needs to be discretized, since it was not defined in Section 2.4, which we rewrite below:

$$
(y_t,v)_\Omega = \int_\Omega y_t v = \frac{d}{dt}\int_\Omega yv.
$$

3.3.1 Discretized form in space

The bilinear form for $(y_t,v)_\Omega$ yields a matrix $\mathbf{M}$ defined as

$$
(\mathbf{M})_{ij} = (\phi_j,\phi_i), \qquad 1\le i,j\le (k+1)N.
$$

Note that $\mathbf{M}$ is a mass matrix, so from (2.26),

$$
\mathbf{M} = -\mathbf{B}.
$$

Recall that $\mathbf{R} = \mathbf{Q} = -\mathbf{B}$ from Section 2.4. If space is one dimensional, the problem is discretized with the same basis functions from Section 2.4 and $\mathbf{M}$ in $\mathbb{R}^{2N}$, defined by

$$\mathbf{M} = \begin{bmatrix} \hat{\mathbf{M}} & & \\ & \ddots & \\ & & \hat{\mathbf{M}} \end{bmatrix}, \qquad \hat{\mathbf{M}} = h \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{bmatrix}.$$

After discretization in space, we get the following problem

$$\min_{(\mathbf{y},\mathbf{u})\in\left(\mathbb{R}^{(k+1)N},\mathbb{R}^{(k+1)N}\right)} \frac{1}{2}\int_0^T (\mathbf{y}(t)-\hat{\mathbf{y}}(t))^T\mathbf{Q}(\mathbf{y}(t)-\hat{\mathbf{y}}(t))dt + \frac{\alpha}{2}\int_0^T \mathbf{u}(t)^T\mathbf{R}\mathbf{u}(t)dt, \tag{3.6}$$

subject to

$$\mathbf{M}\frac{d\mathbf{y}(t)}{dt} + \mathbf{A}\mathbf{y}(t) + \mathbf{B}\mathbf{u}(t) - \mathbf{f}(t) = 0, \tag{3.7}$$

Let $\mathbf{A}$ be defined as

$$\mathbf{A} = \epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}}.$$

Since $\hat{y}$ is given, $\hat{\mathbf{y}}$ can be computed explicitly, which does not depend on $\mathbf{u}$ or $\mathbf{y}$. The value of the objective function changes for different values of $\mathbf{y}$ and $\mathbf{u}$, but $\hat{\mathbf{y}}$ stays constant, thus it can be removed from the objective function to solve an equivalent problem. Solving the discrete optimal control problem defined by (3.6) and (3.7) is equivalent to solving the following discrete optimal control problem:

$$\min_{(\mathbf{y},\mathbf{u})\in\left(\mathbb{R}^{(k+1)N},\mathbb{R}^{(k+1)N}\right)} \frac{1}{2}\int_0^T \left(\mathbf{y}(t)^T\mathbf{Q}(\mathbf{y}(t)) - 2\mathbf{y}(t)^T\hat{\mathbf{y}}\right)dt + \frac{\alpha}{2}\int_0^T \mathbf{u}(t)^T\mathbf{R}\mathbf{u}(t)dt. \tag{3.8}$$

subject to

$$\mathbf{M}\frac{d\mathbf{y}(t)}{dt} + \mathbf{A}\mathbf{y}(t) + \mathbf{B}\mathbf{u}(t) - \mathbf{f}(t) = 0. \tag{3.9}$$

By applying the trapezoid method using K equidistant time steps Δt to the PDE (3.9), we obtain

$$\left(\mathbf{M}+\frac{\Delta t}{2}\mathbf{A}\right)\mathbf{y}(t^{\kappa+1})+\left(\frac{\Delta t}{2}\mathbf{A}-\mathbf{M}\right)\mathbf{y}(t^{\kappa})+\frac{\Delta t}{2}\mathbf{B}\left(\mathbf{u}(t^{\kappa})+\mathbf{u}(t^{\kappa+1})\right) - \frac{\Delta t}{2}\left(\mathbf{f}(t^{\kappa})+\mathbf{f}(t^{\kappa+1})\right)=0, \quad 0\leq\kappa\leq K-1.$$

We approximate $\mathbf{y}(t^{\kappa})$, $\hat{\mathbf{y}}(t^{\kappa})$, $\mathbf{u}(t^{\kappa})$ and $\mathbf{f}(t^{\kappa})$, by constant functions for each time step κ, so we have the following fully discretized set of equations:

$$\left(\mathbf{M}+\frac{\Delta t}{2}\mathbf{A}\right)\mathbf{y}^{\kappa+1}+\left(\frac{\Delta t}{2}\mathbf{A}-\mathbf{M}\right)\mathbf{y}^{\kappa}+\frac{\Delta t}{2}\mathbf{B}\left(\mathbf{u}^{\kappa}+\mathbf{u}^{\kappa+1}\right) - \frac{\Delta t}{2}\left(\mathbf{f}^{\kappa}+\mathbf{f}^{\kappa+1}\right)=0, \quad 0\leq\kappa\leq K-1. \tag{3.10}$$

3.4 Optimization

To solve the optimal control problem, we follow the process outlined in [17]. We can approximate the discretized objective function using the trapezoid rule. Since we are using equidistant time steps, we define Δt^{κ}:

$$\begin{aligned}\Delta t^{\kappa}&=\Delta t, \quad \forall\, 0\leq\kappa\leq K-1,\\ \Delta t^{-1}&=0,\\ \Delta t^{K}&=0.\end{aligned}$$

The objective function (3.8) is then rewritten using the definition of Δt^{κ} and the trapezoid method, to obtain

$$\begin{aligned}&\min_{(\mathbf{y},\mathbf{u})\in\left(\mathbb{R}^{(k+1)N},\mathbb{R}^{(k+1)N}\right)}\frac{1}{2}\int_0^T\left(\mathbf{y}(t)^T\mathbf{Q}\mathbf{y}(t)-2\mathbf{y}(t)^T\hat{\mathbf{y}}(t)dt\right)+\frac{\alpha}{2}\int_0^T\mathbf{u}(t)^T\mathbf{R}\mathbf{u}(t)dt\\ \approx&\min_{(\mathbf{y},\mathbf{u})\in\left(\mathbb{R}^{(k+1)N},\mathbb{R}^{(k+1)N}\right)}\sum_{\kappa=0}^{K-1}\frac{\Delta t^{\kappa}+\Delta t^{\kappa-1}}{2}\left((\mathbf{y}^{\kappa})^T\mathbf{Q}\mathbf{y}^{\kappa}-2(\mathbf{y}^{\kappa})^T\hat{\mathbf{y}}^{\kappa}+\alpha(\mathbf{u}^{\kappa})^T\mathbf{R}\mathbf{u}^{\kappa}\right).\end{aligned} \tag{3.11}$$

By combining the fully discretized PDE (3.10) and the fully discretized objective function (3.11), we obtain the fully discretized optimal control problem:

$$\min_{(\mathbf{y},\mathbf{u})\in\left(\mathbb{R}^{(k+1)N},\mathbb{R}^{(k+1)N}\right)} \sum_{\kappa=0}^{K-1} \frac{\Delta t^{\kappa}+\Delta t^{\kappa-1}}{2}\left((\mathbf{y}^{\kappa})^T\mathbf{Q}\mathbf{y}^{\kappa}-2(\mathbf{y}^{\kappa})^T\hat{\mathbf{y}}^{\kappa}+\alpha(\mathbf{u}^{\kappa})^T\mathbf{R}\mathbf{u}^{\kappa}\right),$$

subject to

$$\left(\mathbf{M}+\frac{\Delta t}{2}\mathbf{A}\right)\mathbf{y}^{\kappa+1}+\left(\frac{\Delta t}{2}\mathbf{A}-\mathbf{M}\right)\mathbf{y}^{\kappa}+\frac{\Delta t}{2}\mathbf{B}\left(\mathbf{u}^{\kappa}+\mathbf{u}^{\kappa+1}\right) - \frac{\Delta t}{2}\left(\mathbf{f}^{\kappa}+\mathbf{f}^{\kappa+1}\right)=0, \quad \forall\, 0\le\kappa\le K-1.$$

3.4.1 Definition of the Lagrangian

Now we introduce the discrete Lagrangian with the Lagrange multiplier $\mathbf{p}$ in $\mathbb{R}^{(k+1)N}$.

$$\begin{aligned} & L(\mathbf{y}^1,...,\mathbf{y}^K,\mathbf{u}^1,...,\mathbf{u}^K,\mathbf{p}^1,...,\mathbf{p}^K) \\ = & \sum_{\kappa=0}^{K-1}\frac{\Delta t^{\kappa}+\Delta t^{\kappa-1}}{2}\left((\mathbf{y}^{\kappa})^T\mathbf{Q}\mathbf{y}^{\kappa}-2(\mathbf{y}^{\kappa})^T\hat{\mathbf{y}}^{\kappa}+\alpha(\mathbf{u}^{\kappa})^T\mathbf{R}\mathbf{u}^{\kappa}\right) \\ + & \sum_{\kappa=0}^{K-1}(\mathbf{p}^{\kappa+1})^T\left(\left(\mathbf{M}+\frac{\Delta t}{2}\mathbf{A}\right)\mathbf{y}^{\kappa+1}+\left(\frac{\Delta t}{2}\mathbf{A}-\mathbf{M}\right)\mathbf{y}^{\kappa}+\frac{\Delta t}{2}\mathbf{B}\left(\mathbf{u}^{\kappa}+\mathbf{u}^{\kappa+1}\right)\right. \\ & \qquad \left.-\frac{\Delta t}{2}\left(\mathbf{f}^{\kappa}+\mathbf{f}^{\kappa+1}\right)\right). \end{aligned}$$

We take the gradient of $L(\mathbf{y},\mathbf{u},\mathbf{p})$ and set it equal to zero to determine the discrete optimality conditions.

$$\nabla_{\mathbf{y}} L(\mathbf{y},\mathbf{u},\mathbf{p}) = \begin{bmatrix} \left(\frac{\Delta t}{2}\mathbf{A}+\mathbf{M}\right)^T \mathbf{p}^1 + \left(\frac{\Delta t}{2}\mathbf{A}-\mathbf{M}\right)^T \mathbf{p}^2 + \Delta t\left(\mathbf{Q}\mathbf{y}^1 - \hat{\mathbf{y}}^1\right) \\ \vdots \\ \left(\frac{\Delta t}{2}\mathbf{A}+\mathbf{M}\right)^T \mathbf{p}^{K-1} + \left(\frac{\Delta t}{2}\mathbf{A}-\mathbf{M}\right)^T \mathbf{p}^K + \Delta t\left(\mathbf{Q}\mathbf{y}^{K-1} - \hat{\mathbf{y}}^{K-1}\right) \\ \left(\frac{\Delta t}{2}\mathbf{A}+\mathbf{M}\right)^T \mathbf{p}^K + \frac{\Delta t}{2}\left(\mathbf{Q}\mathbf{y}^K - \hat{\mathbf{y}}^K\right) \end{bmatrix}$$

$$\nabla_{\mathbf{u}} L(\mathbf{y},\mathbf{u},\mathbf{p}) = \begin{bmatrix} \frac{\Delta t}{2}\alpha\mathbf{R}\mathbf{u}^0 + \frac{\Delta t}{2}\mathbf{B}^T\mathbf{p}^1 \\ \Delta t\alpha\mathbf{R}\mathbf{u}^1 + \frac{\Delta t}{2}\mathbf{B}^T\left(\mathbf{p}^1+\mathbf{p}^2\right) \\ \vdots \\ \Delta t\alpha\mathbf{R}\mathbf{u}^{K-1} + \frac{\Delta t}{2}\mathbf{B}^T\left(\mathbf{p}^{K-1}+\mathbf{p}^K\right) \\ \frac{\Delta t}{2}\alpha\mathbf{R}\mathbf{u}^K + \frac{\Delta t}{2}\mathbf{B}^T\mathbf{p}^K \end{bmatrix}$$

$$\nabla_{\mathbf{p}} L(\mathbf{y},\mathbf{u},\mathbf{p}) = \begin{bmatrix} \left(\frac{\Delta t}{2}\mathbf{A}+\mathbf{M}\right)\mathbf{y}^1 + \left(\frac{\Delta t}{2}\mathbf{A}-\mathbf{M}\right)\mathbf{y}^0 + \frac{\Delta t}{2}\mathbf{B}\left(\mathbf{u}^1+\mathbf{u}^0\right) + \frac{\Delta t}{2}\left(\mathbf{f}^1+\mathbf{f}^0\right) \\ \vdots \\ \left(\frac{\Delta t}{2}\mathbf{A}+\mathbf{M}\right)\mathbf{y}^K + \left(\frac{\Delta t}{2}\mathbf{A}-\mathbf{M}\right)\mathbf{y}^{K-1} + \frac{\Delta t}{2}\mathbf{B}\left(\mathbf{u}^K+\mathbf{u}^{K-1}\right) + \frac{\Delta t}{2}\left(\mathbf{f}^K+\mathbf{f}^{K-1}\right) \end{bmatrix}$$

Let $\mathbf{J}(\mathbf{y},\mathbf{u})$ denote the objective function:

$$\mathbf{J}(\mathbf{y},\mathbf{u}) = \sum_{\kappa=0}^{K-1} \frac{\Delta t^\kappa + \Delta t^{\kappa-1}}{2}\left((\mathbf{y}^\kappa)^T\mathbf{Q}\mathbf{y}^\kappa - 2(\mathbf{y}^\kappa)^T\hat{\mathbf{y}}^\kappa + \alpha(\mathbf{u}^\kappa)^T\mathbf{R}\mathbf{u}^\kappa\right).$$

It can be shown that the gradient of the objective function $\mathbf{J}$ with respect to $\mathbf{u}$ is equal to $\nabla_{\mathbf{u}} L(\mathbf{y},\mathbf{u},\mathbf{p})$ [17]. Since we are minimizing this function, we want to solve for $\mathbf{u}$ such that gradient of the objective function is equal to zero, thus we want to find $\mathbf{u}$ such that

$$\nabla_{\mathbf{u}} L(\mathbf{y},\mathbf{u},\mathbf{p}) = 0.$$

Since this problem has PDE state contraints, $\mathbf{y}$ must solve the state equation and $\mathbf{p}$ must solve the adjoint equation. It can be seen that setting $\nabla_{\mathbf{y}} L(\mathbf{y},\mathbf{u},\mathbf{p})$ equal to zero gives us the discretized adjoint equation and setting $\nabla_{\mathbf{p}} L(\mathbf{y},\mathbf{u},\mathbf{p})$ equal to zero gives us the discretized state equation.

3.4.2 Optimization algorithm

To solve this problem, we implement a Newton Conjugate Gradient Method. This algorithm requires the gradient and the Hessian of the objective function $\mathbf{J}$. From the Lagrangian, we have

$$\nabla_{\mathbf{u}}\mathbf{J}(\mathbf{y},\mathbf{u}) = \nabla_{\mathbf{u}}L(\mathbf{y},\mathbf{u},\mathbf{p}).$$

Since storing the full Hessian could be costly, we only store the vector obtained from multiplying the Hessian to a vector $\mathbf{v}$. We follow the algorithm given in [17] to compute the Hessian-vector multiplication.

Hessian-Vector Multiplication Algorithm

Step 1: Solve the state equation for $\mathbf{y}^{\kappa}$:

$$\left(\mathbf{M}+\frac{\Delta t}{2}\mathbf{A}\right)\mathbf{y}^{\kappa+1}+\left(\frac{\Delta t}{2}\mathbf{A}-\mathbf{M}\right)\mathbf{y}^{\kappa}+\frac{\Delta t}{2}\mathbf{B}\left(\mathbf{u}^{\kappa+1}+\mathbf{u}^{\kappa}\right)+\frac{\Delta t}{2}\left(\mathbf{f}^{\kappa+1}+\mathbf{f}^{\kappa}\right)=0,$$
$$\forall\, 1\le\kappa\le K-1.$$

Step 2: First solve the adjoint equation for $\mathbf{p}^{K}$:

$$\left(\mathbf{M}+\frac{\Delta t}{2}\mathbf{A}\right)^{T}\mathbf{p}^{K}=-\frac{\Delta t}{2}\mathbf{Q}\left(\mathbf{y}^{K}-\hat{\mathbf{y}}^{K}\right),$$

then solve the adjoint for $\mathbf{p}^{\kappa}$:

$$\left(\mathbf{M}+\frac{\Delta t}{2}\mathbf{A}\right)^{T}\mathbf{p}^{\kappa}+\left(\frac{\Delta t}{2}\mathbf{A}-\mathbf{M}\right)^{T}\mathbf{p}^{\kappa+1}=-\Delta t\mathbf{Q}\left(\mathbf{y}^{\kappa}-\hat{\mathbf{y}}^{\kappa}\right),$$
$$\forall\, 1\le\kappa\le K-1.$$

Step 3: First solve for $\mathbf{w}^{1}$:

$$\left(\mathbf{M}+\frac{\Delta t}{2}\mathbf{A}\right)\mathbf{w}^{1}=\frac{\Delta t}{2}\mathbf{B}\left(\mathbf{v}^{1}+\mathbf{v}^{2}\right),$$

then solve for $\mathbf{w}^{\kappa}$:

$$\left(\mathbf{M}+\frac{\Delta t}{2}\mathbf{A}\right)\mathbf{w}^{\kappa}+\left(\frac{\Delta t}{2}\mathbf{A}-\mathbf{M}\right)\mathbf{w}^{\kappa-1}=\frac{\Delta t}{2}\mathbf{B}\left(\mathbf{v}^{\kappa}+\mathbf{v}^{\kappa+1}\right),$$
$$\forall\, 1\leq\kappa\leq K-1.$$

Last solve for $\mathbf{w}^{K}$:

$$\left(\mathbf{M}+\frac{\Delta t}{2}\mathbf{A}\right)\mathbf{w}^{K}+\left(\frac{\Delta t}{2}\mathbf{A}-\mathbf{M}\right)\mathbf{w}^{K-1}=\frac{\Delta t}{2}\mathbf{B}\left(\mathbf{v}^{K}\right).$$

Step 4: First solve for $\mathbf{q}^{K}$:

$$\left(\mathbf{M}+\frac{\Delta t}{2}\mathbf{A}\right)^{T}\mathbf{q}^{K}=\frac{\Delta t}{2}\mathbf{Q}\mathbf{w}^{K},$$

then solve for $\mathbf{q}^{\kappa}$:

$$\left(\mathbf{M}+\frac{\Delta t}{2}\mathbf{A}\right)^{T}\mathbf{q}^{\kappa}+\left(\frac{\Delta t}{2}\mathbf{A}-\mathbf{M}\right)^{T}\mathbf{q}^{\kappa+1}=\Delta t\mathbf{Q}\mathbf{w}^{\kappa},$$
$$\forall\, 1\leq\kappa\leq K-1.$$

Step 5: Compute the multiplication:

$$\nabla^{2}\mathbf{J}(\mathbf{y},\mathbf{u})\mathbf{v}=\begin{bmatrix}\frac{\Delta t}{2}\mathbf{B}^{T}\mathbf{q}^{0}+\alpha\frac{\Delta t}{2}\mathbf{R}\mathbf{v}^{0}\\ \frac{\Delta t}{2}\mathbf{B}^{T}\left(\mathbf{q}^{0}+\mathbf{q}^{1}\right)+\alpha\Delta t\mathbf{R}\mathbf{v}^{1}\\ \vdots\\ \frac{\Delta t}{2}\mathbf{B}^{T}\left(\mathbf{q}^{K-2}+\mathbf{q}^{K-1}\right)+\alpha\Delta t\mathbf{R}\mathbf{v}^{K-1}\\ \frac{\Delta t}{2}\mathbf{B}^{T}\left(\mathbf{q}^{K-1}+\mathbf{q}^{K}\right)+\alpha\frac{\Delta t}{2}\mathbf{R}\mathbf{v}^{K}\end{bmatrix}.$$

Newton Conjugate Gradient Method

We use the Newton Conjugate Gradient method for finding $\mathbf{u}$ that minimizes the objective function. If $\mathbf{u}$ minimizes the objective function, then $\mathbf{u}$ must minimize the following:

$$(\nabla\mathbf{J}(\mathbf{y}(\mathbf{u}),\mathbf{u}),\mathbf{v})+\frac{1}{2}(\nabla^{2}\mathbf{J}(\mathbf{y}(\mathbf{u}),\mathbf{u})\mathbf{v},\mathbf{v}),\quad\forall\mathbf{v}.$$

If we assume

$$(\nabla^2 \mathbf{J}(\mathbf{y}(\mathbf{u}), \mathbf{u})\mathbf{v}, \mathbf{v}) \geq 0, \quad \forall \mathbf{v},$$

and that $\nabla^2 \mathbf{J}(\mathbf{y}(\mathbf{u}), \mathbf{u})$ is invertible, then the unique solution of the minimization problem is the solution $\mathbf{v}$ of

$$\nabla^2 \mathbf{J}(\mathbf{y}(\mathbf{u}), \mathbf{u})\mathbf{v} = -\nabla \mathbf{J}(\mathbf{y}(\mathbf{u}), \mathbf{u}),$$

which is called the Newton step [17]. The Conjugate Gradient method with an Armijo line-search rule is used to solve for $\mathbf{v}$.. The algorithm from [17] is repeated below:

1. Given u_0 and $gtol > 0$. Set $k = 0$, $gtol$ is the tolerance for $||\nabla J(y(u_k), u_k)||$);

2. Compute $\nabla J(y(u_k), u_k)$;

3. If $||\nabla J(y(u_k), u_k)|| < gtol$, stop;

4. Compute $\nabla^2 J(y(u_k), u_k)$;

5. Apply the CG method to compute an approximate solution of the Newton equation $\nabla^2 J(y(u_k), u_k)s_k = -\nabla J(y(u_k), u_k)$;

(i is the iteration index in the CG method);

   5.1. Set $\eta_k \in (0, 1)$, $s_k = 0$ and $p_{k,0} = r_{k,0} = -\nabla J(y(u_k), u_k)$;

   5.2. For $i = 0, 1, ...$ do;

      i. If $||r_{k,i}|| < \eta_k||r_{k,0}||$, go to 5.3;

      ii. Compute $q_{k,i} = \nabla J(y(u_k), u_k)p_i$;

      iii. If $p_{k,i}^T q_{k,i} < 0$, go to 5.3;

      iv. $\gamma_{k,i} = ||r_{k,i}||^2 / p_{k,i}^T q_{k,i}$;

      v. $s_k = s_k + \gamma_{k,i} p_{k,i}$;

      vi. $r_{k,i+1} = r_{k,i} - \gamma_{k,i} q_{k,i}$;

      vii. $\beta_{k,i} = ||r_{k,i+1}||^2 / ||r_{k,i}||^2$;

      viii. $p_{k,i+1} = r_{k,i+1} + \beta_{k,i} p_{k,i}$;

   5.3. If $i = 0$, set $s_k = -\nabla J(y(u_k), u_k)$;

6. Perform Armijo line-search;

   6.1. Set $\alpha_k = 1$ and evaluate $J(y(u_k + \alpha_k s_k), u_k + \alpha_k s_k)$;

   6.2. While

$J(y(u_k + \alpha_k s_k), u_k + \alpha_k s_k) > J(y(u_k), u_k) + 10^{-4}\alpha_k s_k^T \nabla J(y(u_k), u_k)$, do;

      i. Set $\alpha_k = \alpha_k/2$ and evaluate $J(y(u_k + \alpha_k s_k), u_k + \alpha_k s_k)$;

7. Set $u_{k+1} = u_k + \alpha_k s_k$, $k \leftarrow k + 1$. Go to 2.

Algorithm 1: Newton Conjugate Gradient Algorithm with Armijo Line-Search

3.5 Numerical examples

We solve the problem using a software written in Matlab. We use the exact solutions to create the synthetic data to make up $\hat{y}$ and the right hand side f. Recall from the optimality conditions the property that relates u and p:

$$u(x,t) = \frac{p(x,t)}{\alpha}, \quad \forall (x,t) \in (0,1) \times (0,1).$$

For all the examples, we vary the number of intervals the same way and use the following constants:

$$N = 32, 64, 128,$$
$$T = 1,$$
$$\Delta t = \frac{1}{N}.$$

We use the same values for the diffusion coefficient ϵ, the convection coefficient c, and the parameter α for each example:

$$\epsilon = 10^{-9}, \quad c = 1, \quad \alpha = 0.1.$$

Two examples using the trapezoid method for the discretization in time and two examples using backward Euler for the discretization in time are presented. The L^2 error for y_h is defined by $\left(\int_0^1 ||y(t) - y_h(t)||_{L^2(\Omega)} dt \right)^{1/2}$.

Example 1 solves the problem using trapezoid discretization in time and SIPG in space, using linear polynomials, with the following exact solutions:

$$y(x) = \sin(2\pi x)e^{-t},$$

$$p(x) = \sin(\pi x)(1 - t^2),$$

$$u(x) = 10\sin(\pi x)(1 - t^2).$$

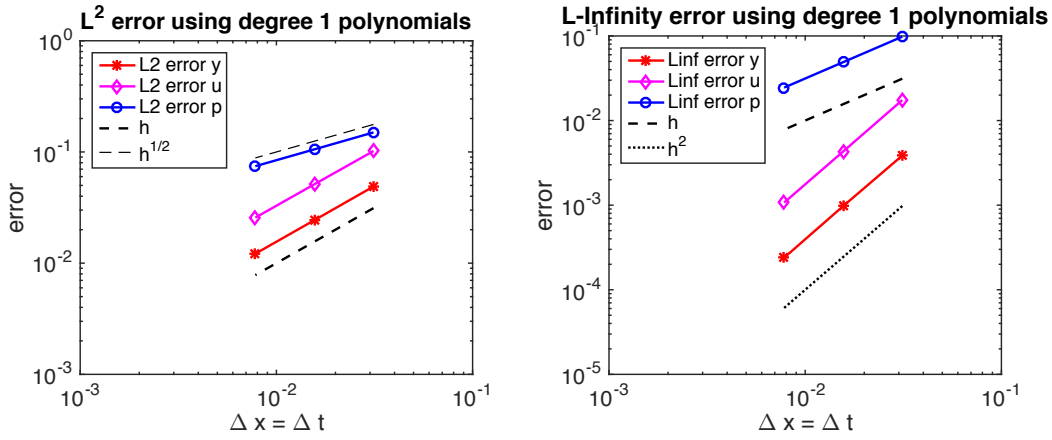


Figure 3.1 : L^2 versus $h = \Delta t$ (left), L^∞ versus $h = \Delta t$ (right) for example 1.

Example 2 solves the problem using trapezoid discretization in time and NIPG in space, using linear polynomials, with the exact solutions from example 1 (Figure 3.1).

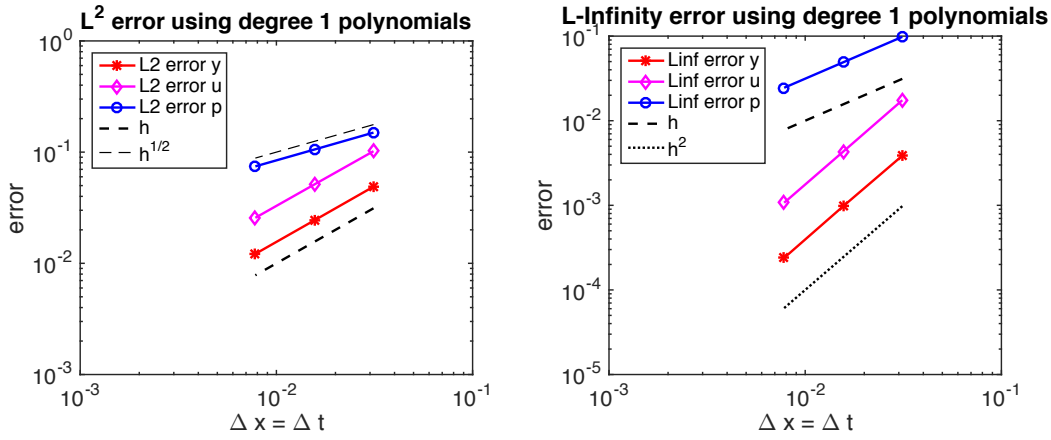


Figure 3.2 : L^2 versus $h = \Delta t$ for example 2.

Example 3 solves the problem using backward Euler discretization in time and SIPG in space, using linear polynomials, with the exact solutions from example 1 (Figure 3.1).

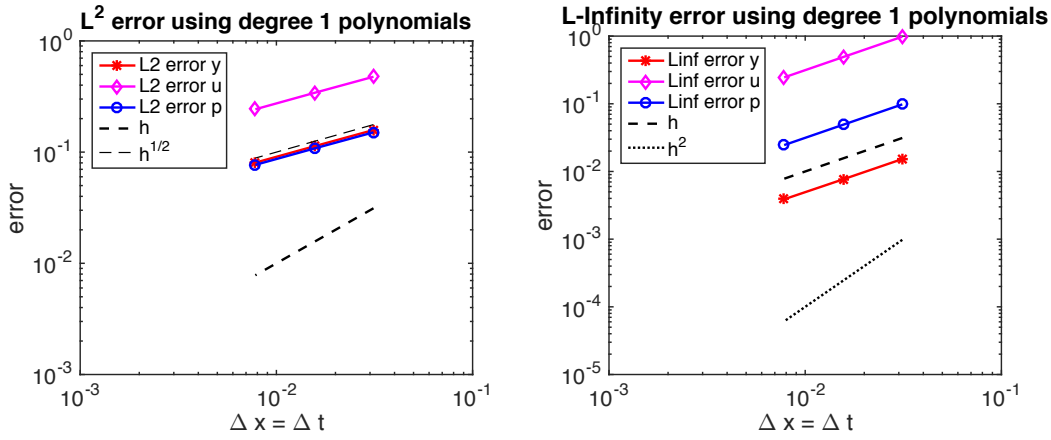


Figure 3.3 : L^2 versus $h = \Delta t$ for example 3.

Example 4 solves the problem using trapezoid discretization in time and SIPG in space, using linear polynomials, with the following exact solutions:

$$y(x) = \cos(x)e^{-t}t^2,$$

$$p(x) = (x^5 - x^4)\sin(\pi t),$$

$$u(x) = 10(x^5 - x^4)\sin(\pi t).$$

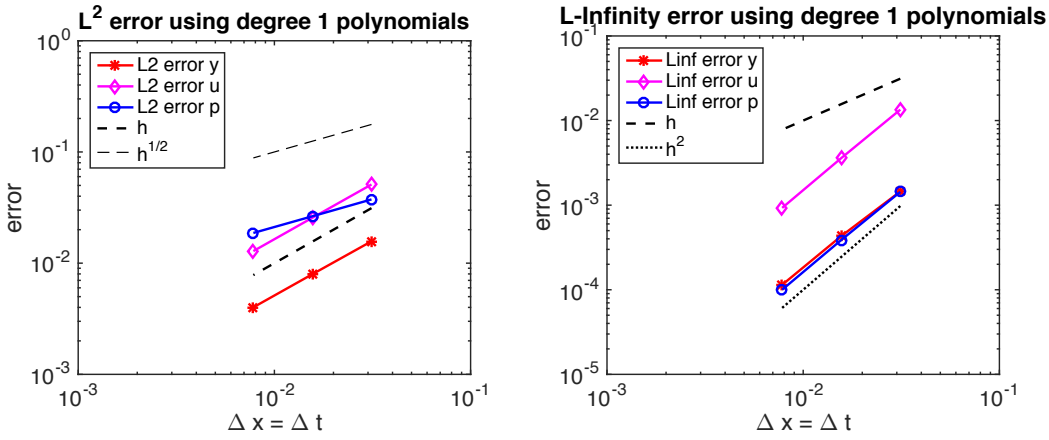


Figure 3.4 : L^2 versus $h = \Delta t$ for example 4.

The numerical results validate the methods we use for solving an optimal control problem governed by a convection-dominated transport equation. We can see that both Backward Euler and trapezoid in time solve the problem successfully and that both SIPG and NIPG work as well. The convergence rates are not as high as seen in the examples in Chapter 2, but the state, control, and adjoint state all converge to the exact solutions.

In Figures 3.1 and 3.4, it is observed that the L^2 error for the state and the control are of order one, but the L^2 error for the adjoint state is of order 1/2. In Figure 3.2, we can see that NIPG works as well as SIPG for this example, since the convergence rates of the variables are the same as in Figure 3.1. When we discretize using backward

Euler in time, the convergence rates are lower than when using trapezoid, which is expected, since backward Euler is a lower order method than the trapezoid method. We can see this in Figure 3.3, which shows that the rates of convergence for the state, control and adjoint are all 1/2.

There has been research showing that the order of the error of the control is not necessarily of the same order as the error of the time stepping method. In [4], the authors prove that when using Runge-Kutta methods of order k for solving time dependent optimal control problems, the L^∞ error of the control is order $k-1$. It can be shown that the Butcher tableaus for backward Euler and trapezoid are as defined below. Backward Euler is an order one Runge-Kutta method, so the error of the control is zero order. Since trapezoid is second order, the control is first order. This loss of convergence rates is due to the fact that when the optimality conditions are derived from the state equation and objective function, the order of the time discretization is applied to the state equation. The time discretization is not directly applied to the adjoint equation, but is determined from discretization of the state equation. Thus the order does not necessarily hold in the time discretization of the adjoint equation. This provides insight as to why the convergence rates are not optimal.

Backward Euler

1	1
	1

Trapezoid

0	0	0
1	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$

3.6 Time Dependent System of PDEs

This next section focuses on solving the optimal control of system of time dependent PDEs. We combine the discretization from Chapters 2.4 and 3.3 to solve the optimal

control problem. Define the state space Y and the control space U to be:

$$Y = \{y \in L^2(0,T;H^1(\Omega)), y_t \in L^2(0,T;H^1(\Omega))\},$$
$$U = L^2(0,T;L^2(\Omega).$$

Let y, z lie in Y and u lie in U. The goal is to solve for y, z, and u in the following optimal control problem:

$$\min_{y,z,u} \frac{1}{2} \int_0^T \int_\Omega \left((y-\hat{y})^2 + (z-\hat{z})^2 + \alpha u^2 \right),$$

subject to

$$\begin{aligned} -\Delta z &= g + u, && \text{in } \Omega \times (0,T], \\ y_t - \epsilon \Delta y + \mathbf{c} \cdot \nabla y &= f + u, && \text{in } \Omega \times (0,T], \\ z &= z_d, && \text{on } \partial\Omega \times (0,T], \\ y &= y_d, && \text{on } \partial\Omega \times (0,T], \\ y &= y_0, && \text{in } \Omega. \end{aligned}$$

The data is the following: $\hat{y}$ and $\hat{z}$ are the desired states, which lie in $L^2(0,T;L^2(\Omega))$, f and g are the data, which also lie in $L^2(0,T;L^2(\Omega))$, $\mathbf{c}$ is a vector in $\mathbb{R}^n$, and $\alpha > 0$.

3.6.1 Discretization

We discretize the PDEs using DG in space, where a_{diff} and a_{conv} define the bilinear forms for the diffusion and convection terms of the PDEs. b is the bilinear operator form used to discretize (u,v) and l_f and l_g are the discretization of the data f and g. The goal is to find v and q in $V_h(\Omega)$ such that

$$\begin{aligned} a_{\text{diff}}(z,q) - b(u,q) - l_g(q) &= 0, \\ (y_t,v) + \epsilon a_{\text{diff}}(y,v) + a_{\text{conv}}(y,v) - b(u,v) - l_f(v) &= 0. \end{aligned}$$

The variational form is rewritten using the explicit matrices defined by the basis functions.

$$\mathbf{A}_{\text{diff}}\mathbf{z}(t) + \mathbf{B}\mathbf{u}(t) - \mathbf{g}(t) = 0,$$
$$\mathbf{M}\frac{d}{dt}\mathbf{y}(t) + (\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}})\mathbf{y}(t) + \mathbf{B}\mathbf{u}(t) - \mathbf{f}(t) = 0.$$

When we discretize the objective function in space, we obtain:

$$\min_{\mathbf{y},\mathbf{z},\mathbf{u}} \frac{1}{2}\int_0^T (\mathbf{y}(t) - \hat{\mathbf{y}}(t))^T\mathbf{Q}(\mathbf{y}(t) - \hat{\mathbf{y}}(t)) + (\mathbf{z}(t) - \hat{\mathbf{z}}(t))^T\mathbf{Q}(\mathbf{z}(t) - \hat{\mathbf{z}}(t)) + \alpha\mathbf{u}(t)^T\mathbf{R}\mathbf{u}(t).$$

We discretize the PDEs in K equal time steps Δt using Crank-Nicolson to obtain:

$$\mathbf{A}_{\text{diff}}\mathbf{z}(t^k) + \mathbf{B}\mathbf{u}(t^k) - \mathbf{g}(t^k) = 0, \quad \forall k = 1, ..., K,$$
$$\left(\mathbf{M} + \frac{\Delta t}{2}(\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}})\right)\mathbf{y}(t^k) + \left(-\mathbf{M} + \frac{\Delta t}{2}(\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}})\right)\mathbf{y}(t^{k-1})$$
$$+ \quad \frac{\Delta t}{2}\mathbf{B}\left(\mathbf{u}(t^k) + \mathbf{u}(t^{k-1})\right) - \frac{\Delta t}{2}\left(\mathbf{f}(t^k) + \mathbf{f}(t^{k-1})\right) = 0, \quad \forall k = 1, ..., K.$$

We discretize the objective function using the trapezoid rule using the same K equal time steps Δt. Let $\Delta t^{-1} = \Delta t^K = 0$ and let $\Delta t^k = \Delta t$ for $k = 0, 1, ..., K-1$.

$$\min_{\mathbf{y},\mathbf{z},\mathbf{u}} \sum_{k=0}^{K} \frac{\Delta t^k + \Delta t^{k-1}}{2}\Bigg((\mathbf{y}(t^k) - \hat{\mathbf{y}}(t^k))^T\mathbf{Q}(\mathbf{y}(t^k) - \hat{\mathbf{y}}(t^k))$$
$$+ (\mathbf{z}(t^k) - \hat{\mathbf{z}}(t^k))^T\mathbf{Q}(\mathbf{z}(t^k) - \hat{\mathbf{z}}(t^k)) + \alpha\mathbf{u}(t^k)^T\mathbf{R}\mathbf{u}(t^k)\Bigg).$$

Let $\mathbf{y}^k = \mathbf{y}(t^k)$ for all k, and similarly for $\mathbf{z}^k$ and $\mathbf{u}^k$. The discretized optimal control problem can be rewritten to be:

$$\min_{\mathbf{y},\mathbf{z},\mathbf{u}} \sum_{k=0}^{K} \frac{\Delta t^k + \Delta t^{k-1}}{2}\left((\mathbf{y}^k - \hat{\mathbf{y}}^k)^T\mathbf{Q}(\mathbf{y}^k - \hat{\mathbf{y}}^k) + (\mathbf{z}^k - \hat{\mathbf{z}}^k)^T\mathbf{Q}(\mathbf{z}^k - \hat{\mathbf{z}}^k) + \alpha(\mathbf{u}^k)^T\mathbf{R}\mathbf{u}^k\right),$$

subject to

$$\mathbf{A}_{\text{diff}}\mathbf{z}^k + \mathbf{B}\mathbf{u}^k - \mathbf{g}^k = 0, \quad \forall k = 1, ..., K,$$
$$\left(\mathbf{M} + \frac{\Delta t}{2}(\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}})\right)\mathbf{y}^k + \left(-\mathbf{M} + \frac{\Delta t}{2}(\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}})\right)\mathbf{y}^{k-1}$$
$$+\frac{\Delta t}{2}\mathbf{B}\left(\mathbf{u}^k + \mathbf{u}^{k-1}\right) - \frac{\Delta t}{2}\left(\mathbf{f}^k + \mathbf{f}^{k-1}\right) = 0, \quad \forall k = 1, ..., K.$$

3.7 Optimization Using the Lagrangian

Introduce adjoint states variables $\mathbf{p}_y$ and $\mathbf{p}_z$ and define the Lagrangian.

$$
\begin{aligned}
L(\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{p}_y,\mathbf{p}_z) = \sum_{k=0}^{K} \frac{\Delta t^k + \Delta t^{k-1}}{2} \Bigg(& (\mathbf{y}^k - \hat{\mathbf{y}}^k)^T \mathbf{Q} (\mathbf{y}^k - \hat{\mathbf{y}}^k) + (\mathbf{z}^k - \hat{\mathbf{z}}^k)^T \mathbf{Q} (\mathbf{z}^k - \hat{\mathbf{z}}^k) \\
& + \alpha (\mathbf{u}^k)^T \mathbf{R} \mathbf{u}^k \Bigg) \\
+ \sum_{k=1}^{K} (\mathbf{p}_y^k)^T \Bigg(& \left(\mathbf{M} + \frac{\Delta t}{2} (\epsilon \mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}}) \right) \mathbf{y}^k \\
& + \left(-\mathbf{M} + \frac{\Delta t}{2} (\epsilon \mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}}) \right) \mathbf{y}^{k-1} \\
& + \frac{\Delta t}{2} \mathbf{B} \left(\mathbf{u}^k + \mathbf{u}^{k-1} \right) - \frac{\Delta t}{2} \left(\mathbf{f}^k + \mathbf{f}^{k-1} \right) \Bigg) \\
+ \sum_{k=1}^{K} (\mathbf{p}_z^k)^T & \left(\mathbf{A}_{\text{diff}} \mathbf{z}^k + \mathbf{B} \mathbf{u}^k - \mathbf{g}^k \right).
\end{aligned}
$$

To determine the discrete optimality conditions, take the gradient of the Lagrangian with respect to $\mathbf{y}$, $\mathbf{z}$, $\mathbf{u}$, $\mathbf{p}_y$, and $\mathbf{p}_z$, and set equal to 0. The optimality conditions are:

$$\begin{bmatrix} \mathbf{Q}(\mathbf{y}^1 - \hat{\mathbf{y}}^1) \\ \vdots \\ \mathbf{Q}(\mathbf{y}^{K-1} - \hat{\mathbf{y}}^{K-1}) \\ \frac{\Delta t}{2}\mathbf{Q}(\mathbf{y}^K - \hat{\mathbf{y}}^K) \end{bmatrix} + \begin{bmatrix} (\mathbf{M} + \frac{\Delta t}{2}(\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}}))^T\mathbf{p}_y^1 + (-\mathbf{M} + \frac{\Delta t}{2}(\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}}))^T\mathbf{p}_y^2 \\ \vdots \\ (\mathbf{M} + \frac{\Delta t}{2}(\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}}))^T\mathbf{p}_y^{K-1} + (-\mathbf{M} + \frac{\Delta t}{2}(\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}}))^T\mathbf{p}_y^K \\ (\mathbf{M} + \frac{\Delta t}{2}(\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}}))^T\mathbf{p}_y^K \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix},$$

$$\begin{bmatrix} \Delta t\mathbf{Q}(\mathbf{z}^1 - \hat{\mathbf{z}}^1) \\ \vdots \\ \Delta t\mathbf{Q}(\mathbf{z}^K - \hat{\mathbf{z}}^K) \end{bmatrix} + \begin{bmatrix} \mathbf{A}_{\text{diff}}^T\mathbf{p}_z^1 \\ \vdots \\ \mathbf{A}_{\text{diff}}^T\mathbf{p}_z^K \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$\begin{bmatrix} \Delta t\alpha\mathbf{R}\mathbf{u}^1 \\ \vdots \\ \Delta t\alpha\mathbf{R}\mathbf{u}^{K-1} \\ \frac{\Delta t}{2}\alpha\mathbf{R}\mathbf{u}^K \end{bmatrix} + \begin{bmatrix} \frac{\Delta t}{2}\mathbf{B}^T(\mathbf{p}_y^2 + \mathbf{p}_y^1) \\ \vdots \\ \frac{\Delta t}{2}\mathbf{B}^T(\mathbf{p}_y^K + \mathbf{p}_y^{K-1}) \\ \frac{\Delta t}{2}\mathbf{B}^T\mathbf{p}_y^K \end{bmatrix} + \begin{bmatrix} \mathbf{B}^T\mathbf{p}_z^1 \\ \vdots \\ \mathbf{B}^T\mathbf{p}_z^K \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$\begin{bmatrix} \mathbf{A}_{\text{diff}}\mathbf{z}^1 \\ \vdots \\ \mathbf{A}_{\text{diff}}\mathbf{z}^K \end{bmatrix} + \begin{bmatrix} \mathbf{B}\mathbf{u}^1 \\ \vdots \\ \mathbf{B}\mathbf{u}^K \end{bmatrix} - \begin{bmatrix} \mathbf{g}^1 \\ \vdots \\ \mathbf{g}^K \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

$$\begin{bmatrix} (\mathbf{M} + \frac{\Delta t}{2}(\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}}))\mathbf{y}^1 + (-\mathbf{M} + \frac{\Delta t}{2}(\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}}))\mathbf{y}^0 \\ \vdots \\ (\mathbf{M} + \frac{\Delta t}{2}(\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}}))\mathbf{y}^K + (-\mathbf{M} + \frac{\Delta t}{2}(\epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}}))\mathbf{y}^{K-1} \end{bmatrix}$$

$$+ \begin{bmatrix} \frac{\Delta t}{2}\mathbf{B}\mathbf{u}^1 \\ \frac{\Delta t}{2}\mathbf{B}(\mathbf{u}^2 + \mathbf{u}^1) \\ \vdots \\ \frac{\Delta t}{2}\mathbf{B}(\mathbf{u}^K + \mathbf{u}^{K-1}) \end{bmatrix} - \begin{bmatrix} \frac{\Delta t}{2}(\mathbf{f}^1 + \mathbf{f}^0) \\ \vdots \\ \frac{\Delta t}{2}(\mathbf{f}^K + \mathbf{f}^{K-1}) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

We can rewrite this in a block 5×5 linear system. For simplicity of notation, let $\tilde{\mathbf{A}} = \epsilon\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{conv}}$. First, define the following blocks:

$$Block11 = \begin{bmatrix} \Delta t\mathbf{Q} & & & \\ & \ddots & & \\ & & \Delta t\mathbf{Q} & \\ & & & \frac{\Delta t}{2}\mathbf{Q} \end{bmatrix},$$

$$Block14 = \begin{bmatrix} (\mathbf{M} + \frac{\Delta t}{2}\tilde{\mathbf{A}})^T & (-\mathbf{M} + \frac{\Delta t}{2}\tilde{\mathbf{A}})^T & & \\ \ddots & \ddots & & \\ & (\mathbf{M} + \frac{\Delta t}{2}\tilde{\mathbf{A}})^T & (-\mathbf{M} + \frac{\Delta t}{2}\tilde{\mathbf{A}})^T & \\ & & (\mathbf{M} + \frac{\Delta t}{2}\tilde{\mathbf{A}})^T & \end{bmatrix},$$

$$Block22 = \begin{bmatrix} \Delta t\mathbf{Q} & & & \\ & \ddots & & \\ & & \Delta t\mathbf{Q} & \\ & & & \frac{\Delta t}{2}\mathbf{Q} \end{bmatrix}, \qquad Block25 = \begin{bmatrix} \mathbf{A}_{\text{diff}}^T & & \\ & \ddots & \\ & & \mathbf{A}_{\text{diff}}^T \end{bmatrix},$$

$$Block33 = \begin{bmatrix} \Delta t\alpha\mathbf{R} & & & \\ & \ddots & & \\ & & \Delta t\alpha\mathbf{R} & \\ & & & \frac{\Delta t}{2}\alpha\mathbf{R} \end{bmatrix}, \quad Block34 = \begin{bmatrix} \frac{\Delta t}{2}\mathbf{B}^T & \frac{\Delta t}{2}\mathbf{B}^T & & \\ & \ddots & \ddots & \\ & & \frac{\Delta t}{2}\mathbf{B}^T & \frac{\Delta t}{2}\mathbf{B}^T \end{bmatrix},$$

$$Block35 = \begin{bmatrix} \mathbf{B}^T & & \\ & \ddots & \\ & & \mathbf{B}^T \end{bmatrix},$$

$$Block41 = \begin{bmatrix} (\mathbf{M} + \frac{\Delta t}{2}\tilde{\mathbf{A}}) & & & \\ (-\mathbf{M} + \frac{\Delta t}{2}\tilde{\mathbf{A}}) & (\mathbf{M} + \frac{\Delta t}{2}\tilde{\mathbf{A}}) & & \\ & \ddots & \ddots & \\ & & (\mathbf{M} + \frac{\Delta t}{2}\tilde{\mathbf{A}}) & (-\mathbf{M} + \frac{\Delta t}{2}\tilde{\mathbf{A}}) \end{bmatrix},$$

$$Block43 = \begin{bmatrix} \frac{\Delta t}{2}\mathbf{B} & \frac{\Delta t}{2}\mathbf{B} & & \\ & \ddots & \ddots & \\ & & \frac{\Delta t}{2}\mathbf{B} & \frac{\Delta t}{2}\mathbf{B} \end{bmatrix},$$

$$Block52 = \begin{bmatrix} \mathbf{A}_{\text{diff}} & & \\ & \ddots & \\ & & \mathbf{A}_{\text{diff}} \end{bmatrix}, \qquad Block53 = \begin{bmatrix} \mathbf{B} & & \\ & \ddots & \\ & & \mathbf{B} \end{bmatrix}$$

Next define the right side vector $\mathbf{b}$:

$$\begin{bmatrix} \Delta t\mathbf{Q}\hat{\mathbf{y}}^1 \\ \vdots \\ \Delta t\mathbf{Q}\hat{\mathbf{y}}^{K-1} \\ \frac{\Delta t}{2}\mathbf{Q}\hat{\mathbf{y}}^K \\ \Delta t\mathbf{Q}\hat{\mathbf{z}}^1 \\ \vdots \\ \Delta t\mathbf{Q}\hat{\mathbf{z}}^K \\ 0 \\ \vdots \\ 0 \\ \frac{\Delta t}{2}(\mathbf{f}^1+\mathbf{f}^0)-(-\mathbf{M}+\frac{\Delta t}{2}\tilde{\mathbf{A}})\mathbf{y}^0 \\ \frac{\Delta t}{2}(\mathbf{f}^2+\mathbf{f}^1) \\ \vdots \\ \frac{\Delta t}{2}(\mathbf{f}^K+\mathbf{f}^{K-1}) \\ \mathbf{g}^1 \\ \vdots \\ \mathbf{g}^K \end{bmatrix}$$

The full block system is:

$$\begin{bmatrix} Block11 & & & Block14 & \\ & Block22 & & & Block25 \\ & & Block33 & Block34 & Block35 \\ Block41 & & Block43 & & \\ & Block52 & Block53 & & \end{bmatrix} \begin{bmatrix} \mathbf{y}=[\mathbf{y}^1,...,\mathbf{y}^K]^T \\ \mathbf{z}=[\mathbf{z}^1,...,\mathbf{z}^K]^T \\ \mathbf{u}=[\mathbf{u}^1,...,\mathbf{u}^K]^T \\ \mathbf{p}_y=[\mathbf{p}_y^1,...,\mathbf{p}_y^K]^T \\ \mathbf{p}_z=[\mathbf{p}_z^1,...,\mathbf{p}_z^K]^T \end{bmatrix} = \mathbf{b},$$

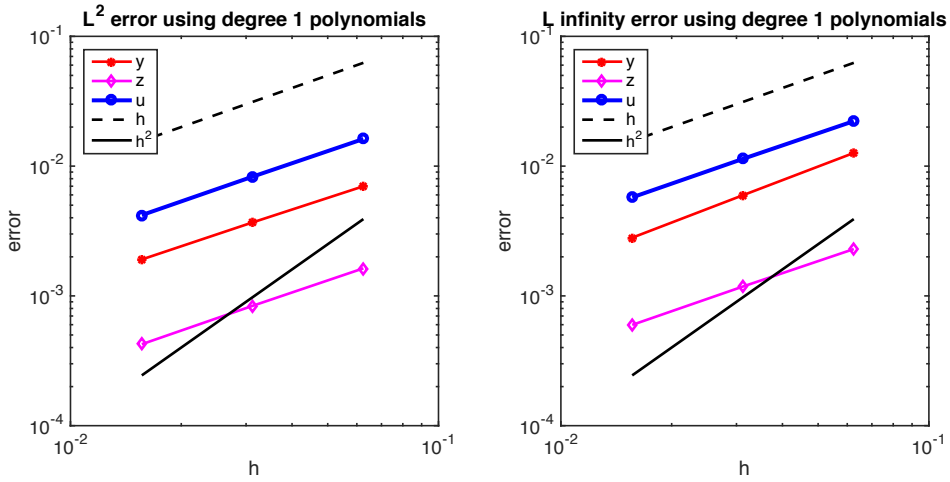
3.7.1 Example

Let $\Omega = (0, 1)$ and $T = 1$. Divide Ω into N equidistant intervals and take K equidistant time steps from 0 to 1. Let $\epsilon = 10^{-9}$, $c = 1$, $\alpha = 1$ and SIPG with $\beta = -1$ and $\sigma_0 = 10$ is used. We approximate the solutions using linear polynomials in space. Let $N = 16, 32, 64$ and for each N, let $\Delta t = 1/N$. Define the following exact solutions:

$$y(x,t) = \cos(t)x^2, \qquad z(x,t) = t^2 e^x$$

$$p_y(x,t) = (x - x^2)(t - 1), \qquad p_z(x,t) = (x - x^2)(t - 1)^2$$

$$u(x,t) = p_y(x,t) + p_z(x,t) = (x - x^2)t(t - 1)$$



The example validates the conclusions from Chapters 2.8 and Chapter 3.5. The approximate solutions to converge to the exact solution for each variable, but they converge with suboptimal rates. Since the trapezoid method was used to discretize the the objective function and PDEs in time, and DG SIPG with linear basis functions in space, the convergence rates are expected to be two, but as seen in Chapter 3.5, the convergence rates for all the variables is one.

Chapter 4

Miscible Displacement Equations

This chapter will give the problem statement of the optimal control of the miscible displacement equations, then derive the optimality conditions. Last, we will show numerical results when solving the PDE without the optimal control problem.

4.1 Miscible displacement equations

The goal is to find the flow rate z in $Z = L^2(0,T)$ such that the corresponding pressure p in P and concentration c in C are as close as possible to the desired pressure $\tilde{p}$ and concentration $\tilde{c}$, both which lie in $L^2(0,T;L^2(\Omega))$.

$$\min_{(c,p,z)\in(C,P,Z)} \frac{1}{2}\int_0^T \int_\Omega \left((c-\tilde{c})^2 + (p-\tilde{p})^2\right) + \frac{1}{2}\int_0^T \alpha z^2,$$

subject to

$$\nabla \cdot \mathbf{u} = z\left(\Phi_I - \Phi_P\right), \qquad \text{in } \Omega \times (0,T), \tag{4.1}$$

$$\mathbf{u} = -K(c)\left(\nabla p - \rho(c)g\right), \qquad \text{in } \Omega \times (0,T), \tag{4.2}$$

$$\phi c_t - \nabla \cdot (D(\mathbf{u})\nabla c) + \mathbf{u}\cdot\nabla c = -z\Phi_I, (c-\hat{c}), \qquad \text{in } \Omega \times (0,T), \tag{4.3}$$

$$q^I = z\Phi_I, \qquad \text{in } \Omega \times (0,T), \tag{4.4}$$

$$q^P = -z\Phi_P, \qquad \text{in } \Omega \times (0,T). \tag{4.5}$$

The boundary conditions and initial condition are:

$$\mathbf{u}(\cdot, t) = 0, \qquad \text{on } \partial\Omega \times (0, T), \tag{4.6}$$

$$D(\mathbf{u})\nabla c \cdot \mathbf{n} = 0, \qquad \text{on } \partial\Omega \times (0, T), \tag{4.7}$$

$$c(\cdot, 0) = c_0(x), \quad \text{in } \Omega, \tag{4.8}$$

where ϕ is the porosity of the porous media, $K(x, c)$ is the absolute permeability of the porous media $\kappa(x)$ divided by the viscosity of the fluid $\mu(c)$, ρ is the density fluid mixture, g is gravity, $D(\mathbf{u})$ is the diffusion dispersion coefficient, $\hat{c}$ is the injected concentration, q^I is the injection source, and q^P is the production sink [13]. Let Φ_I and Φ_P denote the location of the injection and production wells. Note that z is the control, which only depends on time. Assume:

- $K : \Omega \times \mathbb{R} \to \mathbb{R}^{d\times d}$ is symmetric, measurable in x and continuous almost everywhere in c, uniformly bounded and elliptic,
- $D : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is symmetric, Lipzschitz continuous and is defined below:

  $$D(\mathbf{u}) = d_m I + |\mathbf{u}| \left(\alpha_l E(\mathbf{u}) + \alpha_t (I - E(\mathbf{u}))\right),$$

  where α_t is the transverse dispersivity, α_l is the longitudinal dispersivity, d_m is associated with the molecular diffusion, and $E(\mathbf{u})$ is defined as:

  $$E(\mathbf{u}) = \mathbf{u}\mathbf{u}^T |\mathbf{u}|^{-2},$$

  where I is the identity matrix, and $|\cdot|$ is defined to be the Euclidean norm.
- ϕ lies in $L^\infty(\Omega)$ and is bounded above and below by positive constants,
- q^I, q^P lie in $L^\infty(0, T; L^2(\Omega))$, with the property that q^I, q^P are always non-negative, and satisfies

  $$\int_\Omega q^I = \int_\Omega q^P, \qquad \forall t \in [0, T],$$

- ρ maps $\mathbb{R}$ to $\mathbb{R}$ and for some positive constants, ρ is Lipschitz continuous, and bounded above and below by the constants.

4.1.1 Weak form

Define the following spaces:

$$H_0(\Omega; div) = \{v \in L^2(\Omega) : \nabla \cdot v \in L^2(\Omega), v = 0 \text{ in } H^{-1/2}(\Omega)\},$$
$$L^2_0(\Omega) = \{q \in L^2(\Omega) : \int_\Omega q = 0\}.$$

Denote the L^2 inner product on Ω by $(\cdot, \cdot)$. The weak form of the PDE from [13] is given by the following: Find $(\mathbf{u}, p, c, z)$ in $L^\infty[0, T; H_0(\Omega; div)] \times L^\infty[0, T; L^2_0(\Omega)] \times L^2[0, T; H^1(\Omega)] \times L^2(0, T)$ such that

$$\int_0^T \left((K^{-1}(c)\mathbf{u}, \mathbf{v}) - (p, \nabla \cdot \mathbf{v})\right) = \int_0^T (\rho(c)g, \mathbf{v}), \tag{4.9}$$
$$\int_0^T ((q, \nabla \cdot \mathbf{u})) = \int_0^T (z\left(\Phi_I - \Phi_P\right), q), \tag{4.10}$$

for all $(\mathbf{v}, q)$ in $L^2[0, T; H(\Omega; div)] \times L^1[0, T; L^2_0(\Omega)]$ and

$$\int_0^T -(\phi c, w_t) + (D(\mathbf{u})\nabla c, \nabla w) + (\mathbf{u} \cdot \nabla c, w) + (z\Phi_I(c - \hat{c}), w) - (\phi c_0, w(0)) = 0. \tag{4.11}$$

for all w in $\{w \in L^4[0, T; W^{1,4}(\Omega)] \times H^1[0, T; H^1(\Omega)'] : w(T) = 0\}$.

The initial condition is weakly enforced in Ω:

$$\int_\Omega (c(\cdot, 0) - c_0)\, w_0 = 0.$$

4.1.2 Definition of the Lagrangian

The Lagrangian is defined by adding the objective function to the weak form of the PDEs, but the test functions are replaced by adjoint variables that are assumed to lie in the same spaces as their corresponding state variables.

$$
\begin{aligned}
L(\mathbf{u}, p, c, z, \lambda_{\mathbf{u}}, \lambda_p, \lambda_c) = \frac{1}{2}\int_0^T \left((c-\tilde{c}, c-\tilde{c}) + (p-\tilde{p}, p-\tilde{p}) + \alpha z^2\right) \\
+ \int_0^T \left((K^{-1}(c)\mathbf{u}, \lambda_{\mathbf{u}}) - (p, \nabla\cdot\lambda_{\mathbf{u}}) - (\rho(c)g, \lambda_{\mathbf{u}})\right) \\
+ \int_0^T \left((\lambda_p, \nabla\cdot\mathbf{u}) - z(\Phi_I - \Phi_P, \lambda_p)\right) \\
+ \int_0^T \left(-(\phi c, (\lambda_c)_t) + (D(\mathbf{u})\nabla c, \nabla\lambda_c) + (\mathbf{u}\cdot\nabla c, \lambda_c)\right) \\
+ \int_0^T \left(z\Phi_I(c-\hat{c}, \lambda_c) - (\phi c_0, \lambda_c(0))\right) \\
+ \int_\Omega (c(\cdot,0) - c_0)\,\lambda_{c_0}.
\end{aligned}
$$

λ_c is the corresponding adjoint state to c and thus must be in the same space as c. Similarly for $\lambda_{\mathbf{u}}$ and λ_p. For clarification, we note the spaces in which each variable lies:

$$
\begin{aligned}
\mathbf{u}, \lambda_{\mathbf{u}} &\in L^\infty[0,T; H_0(\Omega; div)], \\
p, \lambda_p &\in L^\infty[0,T; L_0^2(\Omega)], \\
c, \lambda_c &\in L^2[0,T; H^1(\Omega)], \\
z &\in L^2(0,T).
\end{aligned}
$$

4.1.3 Derivatives of the Lagrangian

To determine the optimality conditions, the derivative of the Lagrangian is taken with respect to each variable and set equal to zero.

Derivative of L with respect to u

Take the derivative of the Lagrangian with respect to $\mathbf{u}$ and set it equal to zero.

$$
\frac{\partial L}{\partial \mathbf{u}}(\bar{\mathbf{u}}) = 0, \qquad \forall \bar{\mathbf{u}} \in L^\infty[0,T; H_0(\Omega; div)].
$$

Since $D(\mathbf{u})$ maps $\mathbb{R}^d$ to $\mathbb{R}^{d\times d}$, then $D'(\mathbf{u})$ lies in $\mathbb{R}^{d\times d\times d}$ and maps $\mathbb{R}^d$ to $\mathbb{R}^{d\times d}$. Given $\bar{\mathbf{u}}$ in $\mathbb{R}^d$, $D'(\mathbf{u})(\bar{\mathbf{u}})$ lies in $\mathbb{R}^{d\times d}$.

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{u}}(\bar{\mathbf{u}}) &= \int_0^T \Big((K^{-1}(c)\bar{\mathbf{u}}, \lambda_{\mathbf{u}}) + (\lambda_p, \nabla\cdot\bar{\mathbf{u}}) + (D'(\mathbf{u})(\bar{\mathbf{u}})\nabla c, \nabla\lambda_c) + (\bar{\mathbf{u}}, \nabla\lambda_c)\Big) \\
&= \int_0^T \Big((K^{-1}(c)\bar{\mathbf{u}}, \lambda_{\mathbf{u}}) + (D'(\mathbf{u})(\bar{\mathbf{u}})\nabla c, \nabla\lambda_c) + (\bar{\mathbf{u}}, \nabla\lambda_c) + (\lambda_p, \bar{\mathbf{u}}\cdot\mathbf{n})_{\partial\Omega} - (\bar{\mathbf{u}}, \nabla\lambda_p)\Big) \\
&= \int_0^T \Big((K^{-1}(c)\bar{\mathbf{u}}, \lambda_{\mathbf{u}}) + (D'(\mathbf{u})(\bar{\mathbf{u}})\nabla c, \nabla\lambda_c) + (\bar{\mathbf{u}}, \nabla\lambda_c) - (\bar{\mathbf{u}}, \nabla\lambda_p)\Big) \\
&= \int_0^T \Big((\bar{\mathbf{u}}, K^{-1}(c)^T\lambda_{\mathbf{u}}) + (D'(\mathbf{u})(\bar{\mathbf{u}})\nabla c, \nabla\lambda_c) + (\bar{\mathbf{u}}, \nabla\lambda_c - \nabla\lambda_p)\Big) \\
&= \int_0^T \Big((D'(\mathbf{u})(\bar{\mathbf{u}})\nabla c, \nabla\lambda_c) + (\bar{\mathbf{u}}, (K^{-1}(c))^T\lambda_{\mathbf{u}} + \nabla\lambda_c - \nabla\lambda_p)\Big)
\end{aligned}$$

To determine the equation derived from setting $\frac{\partial L}{\partial \mathbf{u}}(\bar{\mathbf{u}})$ equal to zero, $\bar{\mathbf{u}}$ needs to be isolated. The following lemma, which is proven in the Appendix, is used to do this.

Lemma 4.1

$$(D'(\mathbf{u})(\bar{\mathbf{u}})\nabla c)\cdot\nabla\lambda_c = \bar{\mathbf{u}}\cdot F(\mathbf{u}, \nabla c, \nabla\lambda_c),$$

where we define F by

$$F(\mathbf{u}, \nabla c, \nabla\lambda_c) = \frac{\alpha_l - \alpha_t}{|\mathbf{u}|}\left(\nabla\lambda_c\nabla c^T\mathbf{u} + \nabla c\nabla\lambda_c^T\mathbf{u}\right) + \mathbf{u}\nabla\lambda_c^T\left(\frac{(\alpha_t - \alpha_l)E(\mathbf{u}) + \alpha_t I}{|\mathbf{u}|}\right)\nabla c.$$

Using Lemma 4.1, we have

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{u}}(\bar{\mathbf{u}}) &= \int_0^T \Big((D'(\mathbf{u})(\bar{\mathbf{u}})\nabla c, \nabla\lambda_c) + (\bar{\mathbf{u}}, (K^{-1}(c))^T\lambda_{\mathbf{u}} + \nabla\lambda_c - \nabla\lambda_p)\Big) \\
&= \int_0^T \Big((\bar{\mathbf{u}}, F(\mathbf{u}, \nabla c, \nabla\lambda_c)) + (\bar{\mathbf{u}}, (K^{-1}(c))^T\lambda_{\mathbf{u}} + \nabla\lambda_c - \nabla\lambda_p)\Big)
\end{aligned}$$

Since this holds for all $\bar{\mathbf{u}}$, this gives the adjoint state equation with respect to λ_p:

$$K^{-1}(c)^T\lambda_{\mathbf{u}} + F(\mathbf{u}, \nabla c, \nabla\lambda_c) + \nabla\lambda_c - \nabla\lambda_p = 0 \qquad \text{in } \Omega\times(0,T). \tag{4.12}$$

Derivative of L with respect to p

Next, take the derivative of the Lagrangian with respect to p and set it equal to zero.

$$\frac{\partial L}{\partial p}(\bar{p}) = 0, \qquad \forall \bar{p} \in L^{\infty}[0,T;L_0^2(\Omega)].$$

$$\begin{aligned}\frac{\partial L}{\partial p}(\bar{p}) &= \int_0^T \left((p-\tilde{p},\bar{p}) - (\bar{p},\nabla\cdot\lambda_{\mathbf{u}})\right) \\ &= \int_0^T (p-\tilde{p}-\nabla\cdot\lambda_{\mathbf{u}},\bar{p}) \\ &= 0.\end{aligned}$$

Since this holds for all $\bar{p}$, this gives us the adjoint state equation with respect to $\lambda_{\mathbf{u}}$:

$$(p-\tilde{p}) - \nabla\cdot\lambda_{\mathbf{u}} = 0, \qquad \text{in } \Omega\times(0,T). \tag{4.13}$$

Derivative of L with respect to c

Then, take the derivative of the Lagrangian with respect to c and set it equal to zero.

$$\frac{\partial L}{\partial c}(\bar{c}) = 0, \qquad \forall \bar{c} \in L^2[0,T;H^1(\Omega)].$$

$$\begin{aligned}\frac{\partial L}{\partial c}(\bar{c}) = \int_0^T \Bigg(&(c-\tilde{c},\bar{c}) + \left(\left(\frac{\partial}{\partial c}K^{-1}(c)\bar{c}\right)\mathbf{u},\lambda_{\mathbf{u}}\right) - (\rho'(c)\bar{c}g,\lambda_{\mathbf{u}}) - (\phi\bar{c},(\lambda_c)_t) \\ &+ (D(\mathbf{u})\nabla\bar{c},\nabla\lambda_c) + (\lambda_c\mathbf{u},\nabla\bar{c}) + z(\bar{c},\lambda_c)\Bigg).\end{aligned} \tag{4.14}$$

We rewrite K is rewritten as a function of the permeability $\kappa(x)$ and the viscosity $\mu(c)$ to obtain K^{-1}:

$$K^{-1}(c) = \kappa^{-1}(x)\mu(c).$$

$\frac{\partial}{\partial c}K^{-1}(c)$ is rewritten as

$$\begin{aligned}\frac{\partial}{\partial c}K^{-1}(c) &= \frac{\partial}{\partial c}\left(\kappa^{-1}(x)\mu(c)\right)\\ &= \kappa^{-1}(x)\mu'(c). \end{aligned} \tag{4.15}$$

Next, rewrite (4.14) using (4.15):

$$\begin{aligned}\frac{\partial L}{\partial c}(\bar{c}) &= \int_0^T \Bigg((c-\tilde{c},\bar{c}) + \left(\left(\kappa^{-1}\mu'(c)\bar{c}\right)\mathbf{u},\lambda_{\mathbf{u}}\right) - (\rho'(c)(g\cdot\lambda_{\mathbf{u}}),\bar{c}) - (\phi(\lambda_c)_t,\bar{c})\\ &\quad + z(\lambda_c,\bar{c}) + (D(\mathbf{u})\nabla\bar{c},\nabla\lambda_c) + (\lambda_c\mathbf{u},\nabla\bar{c})\Bigg)\\ &= \int_0^T \Bigg((c-\tilde{c},\bar{c}) + \left(\kappa^{-1}\mu'(c)(\mathbf{u}\cdot\lambda_{\mathbf{u}}),\bar{c}\right) - (\rho'(c)(g\cdot\lambda_{\mathbf{u}}),\bar{c}) - (\phi(\lambda_c)_t,\bar{c})\\ &\quad + z(\lambda_c,\bar{c}) + (\nabla\bar{c},D(\mathbf{u})^T\nabla\lambda_c) + (\nabla\bar{c},\lambda_c\mathbf{u})\Bigg)\\ &= \int_0^T \Bigg((c-\tilde{c},\bar{c}) + \left(\kappa^{-1}\mu'(c)(\mathbf{u}\cdot\lambda_{\mathbf{u}}),\bar{c}\right) - (\rho'(c)(g\cdot\lambda_{\mathbf{u}}),\bar{c}) - (\phi(\lambda_c)_t,\bar{c}) + z(\lambda_c,\bar{c})\\ &\quad + (D(\mathbf{u})^T\nabla\lambda_c\cdot\mathbf{n},\bar{c})_{\partial\Omega} - (\nabla\cdot(D(\mathbf{u})^T\nabla\lambda_c),\bar{c}) + (\lambda_c\mathbf{u}\cdot\mathbf{n},\bar{c})_{\partial\Omega} + (\nabla\cdot(\lambda_c\mathbf{u}),\bar{c})\Bigg)\\ &= \int_0^T \Bigg((c-\tilde{c},\bar{c}) + \left(\kappa^{-1}\mu'(c)(\mathbf{u}\cdot\lambda_{\mathbf{u}}),\bar{c}\right) - (\rho'(c)(g\cdot\lambda_{\mathbf{u}}),\bar{c}) - (\phi(\lambda_c)_t,\bar{c}) + z(\lambda_c,\bar{c})\\ &\quad + (D(\mathbf{u})^T\nabla\lambda_c\cdot\mathbf{n},\bar{c})_{\partial\Omega} - (\nabla\cdot(D(\mathbf{u})^T\nabla\lambda_c),\bar{c}) + (\nabla\cdot(\lambda_c\mathbf{u}),\bar{c})\Bigg)\\ &= 0. \end{aligned}$$

We set $(D(\mathbf{u})^T\nabla\lambda_c)\cdot\mathbf{n}$ equal to zero on the boundary.

$$(D(\mathbf{u})^T\nabla\lambda_c)\cdot\mathbf{n} = 0, \qquad \text{on } \partial\Omega\times(0,T).$$

Next, we rewrite $\kappa^{-1}\mu'(c)$ from (4.15) and then for all $\bar{c}$,

$$\begin{aligned}&\int_0^T \Bigg((c-\tilde{c},\bar{c}) + \left(\frac{\partial}{\partial c}K^{-1}(c)(\mathbf{u}\cdot\lambda_{\mathbf{u}}),\bar{c}\right) - (\rho'(c)(g\cdot\lambda_{\mathbf{u}}),\bar{c}) - (\phi(\lambda_c)_t,\bar{c}) + (z\lambda_c,\bar{c})\\ &\quad - (\nabla\cdot(D(\mathbf{u})^T\nabla\lambda_c),\bar{c}) + (\nabla\cdot(\lambda_c\mathbf{u}),\bar{c})\Bigg) = 0.\end{aligned}$$

Since this holds for all $\bar{c}$, the adjoint state equation with respect to λ_c is obtained.

$$(c-\tilde{c}) + \frac{\partial}{\partial c}K^{-1}(c)(\mathbf{u}\cdot\lambda_{\mathbf{u}}) - \rho'(c)(g\cdot\lambda_{\mathbf{u}}) - \phi(\lambda_c)_t - \nabla\cdot(D(\mathbf{u})^T\nabla\lambda_c + \lambda_c\mathbf{u}) + z\lambda_c = 0,$$

in $\Omega\times(0,T)$, with the following boundary conditions:

$$\left(D(\mathbf{u})^T\nabla\lambda_c\right)\cdot\mathbf{n} = 0, \qquad \text{on } \partial\Omega\times(0,T).$$

Derivative of L with respect to z

Take the derivative of the Lagrangian with respect to z and set it equal to zero.

$$\frac{\partial L}{\partial z}(\bar{z}) = 0, \qquad \forall\bar{z}\in L^2(0,T).$$

$$\begin{aligned}\frac{\partial L}{\partial z}(\bar{z}) &= \int_0^T (\alpha z\bar{z} - \bar{z}(\Phi_I-\Phi_P,\lambda_p) + \bar{z}(\Phi_I(c-\hat{c}),\lambda_c)) \\ &= \int_0^T (\bar{z}\left(\alpha z - (\Phi_I-\Phi_P,\lambda_p) + (\Phi_I(c-\hat{c}),\lambda_c)\right)) = 0.\end{aligned}$$

This gives us the following equation:

$$\alpha z - (\Phi_I-\Phi_P,\lambda_p) + (\Phi_I(c-\hat{c}),\lambda_c) = 0, \qquad \text{in } (0,T). \tag{4.16}$$

Derivative of L with respect to λ_p

Next, take the derivative of the Lagrangian with respect to λ_p and set it equal to zero.

$$\frac{\partial L}{\partial \lambda_p}(\bar{\lambda}) = 0, \qquad \forall\bar{\lambda}\in L^\infty[0,T;L_0^2(\Omega)].$$

$$\begin{aligned}\frac{\partial L}{\partial \lambda_p}(\bar{\lambda}) &= \int_0^T \left((\bar{\lambda}, \nabla \cdot \mathbf{u}) - z(\Phi_I - \Phi_P, \bar{\lambda})\right) \\ &= \int_0^T \left((\nabla \cdot \mathbf{u}) - z(\Phi_I - \Phi_P), \bar{\lambda}\right) \\ &= 0.\end{aligned}$$

Since this holds for all $\bar{\lambda}$, this recovers the state equation with respect to $\mathbf{u}$.

$$\nabla \cdot \mathbf{u} = z(\Phi_I - \Phi_P), \qquad \text{in } \Omega \times (0, T). \tag{4.17}$$

Derivative of L with respect to $\lambda_{\mathbf{u}}$

Then, take the derivative of the Lagrangian with respect to $\lambda_{\mathbf{u}}$ and set it equal to zero.

$$\frac{\partial L}{\partial \lambda_{\mathbf{u}}}(\bar{\lambda}) = 0, \qquad \forall \bar{\lambda} \in L^\infty[0, T; H_0(\Omega; div)].$$

$$\begin{aligned}\frac{\partial L}{\partial \lambda_{\mathbf{u}}}(\bar{\lambda}) &= \int_0^T \left((K^{-1}(c)\mathbf{u}, \bar{\lambda}) - (p, \nabla \cdot \bar{\lambda}) - (\rho(c)g, \bar{\lambda})\right) \\ &= \int_0^T \left((K^{-1}(c)\mathbf{u}, \bar{\lambda}) - (p, \bar{\lambda} \cdot \mathbf{n})_{\partial\Omega} + (\bar{\lambda}, \nabla p) - (\rho(c)g, \bar{\lambda})\right) \\ &= \int_0^T \left((K^{-1}(c)\mathbf{u}, \bar{\lambda}) + (\bar{\lambda}, \nabla p) - (\rho(c)g, \bar{\lambda})\right) \\ &= \int_0^T \left((K^{-1}(c)\mathbf{u} + \nabla p - \rho(c)g, \bar{\lambda})\right) \\ &= 0.\end{aligned}$$

Since this holds for all $\bar{\lambda}$, this recovers the state equation with respect to $\mathbf{u}, p$.

$$K^{-1}(c)\mathbf{u} + \nabla p - \rho(c)g = 0, \qquad \text{in } \Omega \times (0, T). \tag{4.18}$$

Combine (4.17) and (4.18) to get the PDE of Darcy's Law.

$$\mathbf{u} = -K(c)\left(\nabla p - \rho(c)g\right), \qquad \text{in } \Omega \times (0, T). \tag{4.19}$$

Derivative of L with respect to λ_c

Take the derivative of the Lagrangian with respect to λ_c and set it equal to zero.

$$\frac{\partial L}{\partial \lambda_c}(\bar{\lambda}) = 0, \qquad \forall \bar{\lambda} \in L^2[0,T;H^1(\Omega)].$$

$$\begin{aligned}
\frac{\partial L}{\partial \lambda_c}(\bar{\lambda}) &= \int_0^T \big(-(\phi c, \bar{\lambda}_t) + (D(\mathbf{u})\nabla c, \nabla\bar{\lambda}) + (\mathbf{u}\cdot\nabla c, \bar{\lambda}) + z(\Phi_I(c-\hat{c}), \bar{\lambda})\big) - (\phi c_0, \bar{\lambda}(0)) \\
&= \int_\Omega \big(-\phi(\cdot)c(\cdot,T)\bar{\lambda}(\cdot,T) + \phi(\cdot)c(\cdot,0)\bar{\lambda}(\cdot,0)\big) \\
&+ \int_0^T \big((\phi c_t, \bar{\lambda}) + (D(\mathbf{u})\nabla c, \bar{\lambda})_{\partial\Omega} - (\nabla\cdot(D(\mathbf{u})\nabla c), \bar{\lambda}) + (\mathbf{u}\cdot\nabla c, \bar{\lambda}) + z(\Phi_I(c-\hat{c}), \bar{\lambda})\big) \\
&- (\phi c_0, \bar{\lambda}(0)) \\
&= -\int_\Omega \phi(\cdot)c(\cdot,T)\bar{\lambda}(\cdot,T) \\
&+ \int_0^T \big((\phi c_t, \bar{\lambda}) - (\nabla\cdot(D(\mathbf{u})\nabla c), \bar{\lambda}) + (\mathbf{u}\cdot\nabla c, \bar{\lambda}) + z(\Phi_I(c-\hat{c}), \bar{\lambda})\big) \\
&= 0.
\end{aligned}$$

We choose $\lambda_c(\cdot,T)$ to vanish in Ω, so that for all λ in $L^2[0,T;H^1(\Omega)]$:

$$\int_0^T \big((\phi c_t, \bar{\lambda}) - (\nabla\cdot(D(\mathbf{u})\nabla c), \bar{\lambda}) + (\mathbf{u}\cdot\nabla c, \bar{\lambda}) + z(\Phi_I(c-\hat{c}), \bar{\lambda})\big) = 0.$$

This recovers the state equation with respect to c, the transport equation.

$$\phi c_t - \nabla\cdot(D(\mathbf{u})\nabla c) + \mathbf{u}\cdot\nabla c + z\Phi_I(c-\hat{c}) = 0, \qquad \text{in } \Omega\times(0,T).$$

This also recovers a condition on λ_c:

$$\lambda_c(\cdot,T) = 0, \qquad \text{in } \Omega.$$

Derivative of L with respect to λ_{c_0}

Finally, take the derivative of the Lagrangian with respect to λ_{c_0} and set it equal to zero.

$$\frac{\partial L}{\partial \lambda_{c_0}}(\bar{\lambda}) = 0, \qquad \forall \bar{\lambda} \in H^1(\Omega).$$

$$\frac{\partial L}{\partial \lambda_{c_0}}(\bar{\lambda}) = (c(\cdot,0) - c_0, \lambda_{c_0}) = 0.$$

This recovers the initial condition for the state equation with respect to c.

$$c(\cdot,0) = c_0, \qquad \text{in } \Omega. \tag{4.20}$$

4.1.4 Optimality conditions

The optimality conditions derived in the previous section are summarized below.

$$\begin{aligned}
(p - \tilde{p}) - \nabla \cdot \lambda_{\mathbf{u}} &= 0, && \text{in } \Omega \times (0,T),\\
-\phi(\lambda_c)_t - \nabla \cdot \left(D(\mathbf{u})^T \nabla \lambda_c + \mathbf{u}\lambda_c\right) + z\lambda_c &= \tilde{c} - c - \left(\frac{\partial}{\partial c} K^{-1}(c)(\mathbf{u}\cdot\lambda_{\mathbf{u}}) + \rho'(c) g \cdot \lambda_{\mathbf{u}}\right), \\
& && \text{in } \Omega \times (0,T),\\
K^{-1}(c)\lambda_{\mathbf{u}} - \nabla\lambda_p + F(\mathbf{u}, \nabla c, \nabla\lambda_c) + \nabla\lambda_c &= 0, && \text{in } \Omega \times (0,T),\\
\nabla \cdot \mathbf{u} &= z(\Phi_I - \Phi_P), && \text{in } \Omega \times (0,T),\\
\mathbf{u} &= -K(c)\left(\nabla p - \rho(c) g\right), && \text{in } \Omega \times (0,T),\\
\phi c_t - \nabla \cdot (D(\mathbf{u})\nabla c) + \mathbf{u}\cdot\nabla c + z\Phi_I(c - \hat{c}) &= 0, && \text{in } \Omega \times (0,T),\\
\alpha z - (\Phi_I - \Phi_P, \lambda_p) + (\Phi_I(c - \hat{c}), \lambda_c) &= 0, && \text{in } (0,T),\\
\mathbf{u} &= 0, && \text{on } \partial\Omega \times (0,T),\\
D(\mathbf{u})^T \nabla \lambda_c &= 0, && \text{on } \partial\Omega \times (0,T),\\
D(\mathbf{u})\nabla c &= 0, && \text{on } \partial\Omega \times (0,T),\\
\lambda_c(\cdot,T) &= 0, && \text{in } \Omega,\\
c(\cdot,0) &= c_0, && \text{in } \Omega.
\end{aligned}$$

4.2 Discretization

We discretize the miscible displacement equations using discontinuous Galerkin methods in both time and space. We rewrite the PDEs without z, Φ_I and Φ_P, since we will only solve the system of PDEs. Following the discretization given in [13], DG1 in space, where the basis polynomials are degree 1 is used, as well as DG0 and DG1 in time, where the basis polynomials are degree 0 and 1 respectively. Note that DG0 is equivalent to modified Backward Euler and DG1 is equivalent to modified Crank-Nicolson.

The goal is to solve the following system of PDEs, which is equivalent to (4.1)-(4.8)

$$\begin{aligned} \nabla \cdot \mathbf{u} &= q^I - q^P, && \text{in } \Omega \times (0,T], \\ \mathbf{u} &= -K(c)\left(\nabla p - \rho(c) g\right), && \text{in } \Omega \times (0,T], \\ \phi c_t - \nabla \cdot (D(\mathbf{u})\nabla c) + \mathbf{u} \cdot \nabla c + q^I c &= \hat{c} q^I, && \text{in } \Omega \times (0,T], \end{aligned}$$

with following boundary conditions and initial condition:

$$\begin{aligned} \mathbf{u}(\cdot, t) &= 0, && \text{on } \partial\Omega \times (0,T) \\ D(\mathbf{u})\nabla c &= 0, && \text{on } \partial\Omega \times (0,T) \\ c(\cdot, 0) &= c_0(x), && \text{in } \Omega \end{aligned}$$

Let $P_k(E)$ be the set of all polynomials of degree k in the element E. Define the following spaces, where E_h is the mesh over Ω and Γ_h is the set of faces in the mesh.

$$\begin{aligned} U_h &= \{u \in H(\Omega; div) | u|_E \in P_k(E) + xP_k(E), E \in E_h\}, \\ P_h &= \{q_h \in L^2(\Omega) : q_h | E \in P_k(E), E \in E_h\}, \\ C_h &= \{c_h \in H^1(E_h) : q_h | E \in P_l(E), E \in E_h\}. \end{aligned}$$

The numerical scheme is defined as follows:

$$\int_{t_{k-1}}^{t_k} \left(\left(K^{-1}(c_h)\mathbf{u}_h, \mathbf{v}_h \right) - (p_h, \nabla \cdot \mathbf{v}_h) \right) = \int_{t_{k-1}}^{t_k} (\rho(c_h)g, \mathbf{v}_h),$$

$$\int_{t_{k-1}}^{t_k} (q_h, \nabla \cdot \mathbf{u}_h) = \int_{t_{k-1}}^{t_k} \left((q^I - q^P), q_h \right),$$

$$\int_{t_{k-1}}^{t_k} ((\phi(c_h)_t, w_h) + B_d(c_h, w_h; \mathbf{u}_h) + B_{cq}(c_h, w_h; \mathbf{u}_h)) + ([c_h^{k-1}]_t, \phi w_{h+}^{k-1}) = \int_{t_{k-1}}^{t_k} \left(\hat{c}q^T, w_h \right),$$

for all $\mathbf{v}_h$ in $P_l[t^{k-1}, t^k; U_h]$, q_h in $P_l[t^{k-1}, t^k; P_h]$, and w_h in $P_l[t^{k-1}, t^k; C_h]$.

Define B_d and B_{cq} as the discretizations of $-\nabla \cdot (D(\mathbf{u})\nabla c)$ and $-\mathbf{u}\nabla c + q^I c$, respectively.

$$\begin{aligned} B_d(c_h, w_h; \mathbf{u}_h) &= (D(\mathbf{u})\nabla c_h, \nabla w_h)_{E_h} - ([w_h], \{D(\mathbf{u}_h)\nabla c_h\})_{\Gamma_h} \\ &\quad + \beta([c_h], \{D(\mathbf{u}_h)\nabla w_h\})_{\Gamma_h} + \frac{\sigma_0}{h}(1 + \{|\mathbf{u}_h|\})[c_h], [w_h])_{\Gamma_h}, \\ B_{cq}(c_h, w_h; \mathbf{u}_h) &= \frac{1}{2} \left((\mathbf{u}_h \nabla c_h, w_h)_{E_h} - (\mathbf{u}_h c_h, \nabla w_h)_{E_h} + \left((q^I - q^P)c_h, w_h \right) \right. \\ &\quad \left. + (c_h^{up}\mathbf{u}_h, [w_h])_{\Gamma_h} - (w_h^{down}\mathbf{u}_h, [c_h])_{\Gamma_h} \right). \end{aligned}$$

4.3 Numerical results

In this section, we discuss two sets of numerical examples when solving the miscible displacement equations without minimizing the objective function. In [13], error estimates are given for H^1 and L^2 errors for the pressure and the concentration.

4.3.1 One dimension

We solve the problem using a software in Matlab. The exact solutions are used to create the synthetic data to make up the initial and known data. For all the examples, we vary the number of intervals N the same way:

$$N = 8, 16, 32, 64.$$

We use linear polynomials in space and SIPG with the penalty parameter σ_0. We choose the final time T:

$$\sigma_0 = 10, \qquad T = 1.$$

For each example, we use the following data and exact solutions:

$$g(x,t) = 0,$$

$$q^I(x,t) = 0, \qquad q^P(x,t) = 0,$$

$$\phi(x) = 1, \qquad \hat{c}(x,t) = 0,$$

$$D(u(x)) = \sqrt{x^2+1}, \qquad K(c) = 10^{-4}(.5c + .18(1-c))^4,$$

$$c(x,t) = 1 + cos(\pi t^2)e^{-x^2}, \qquad p(x,t) = cos(\pi x)(t^2+1).$$

Example 1 solves the problem using DG0 in time with $\Delta t = \frac{1}{N^2}$.

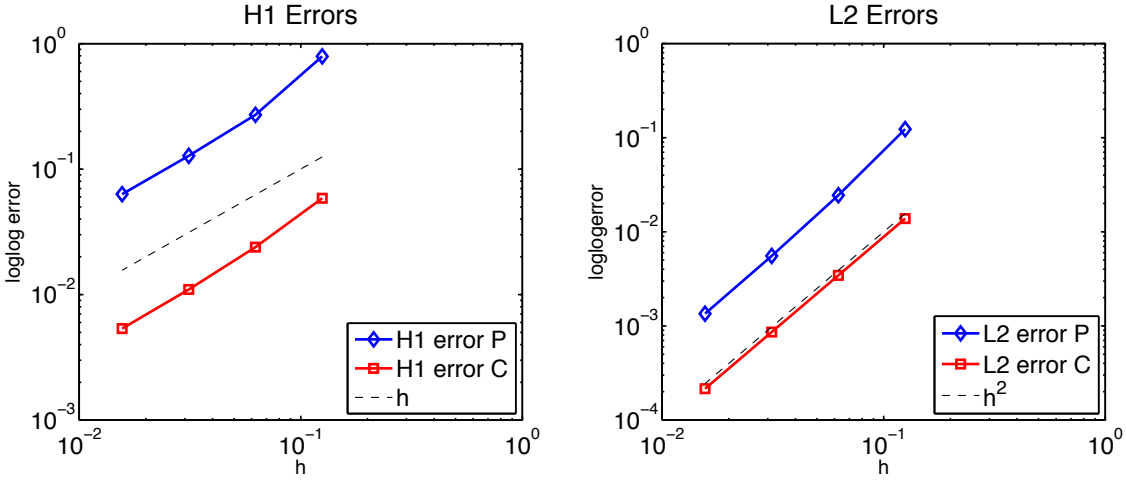


Figure 4.1 : H^1 versus h (left), L^2 versus h (right) for example 1.

Example 2 solves the problem using DG1 in time with $\Delta t = \frac{1}{N}$.

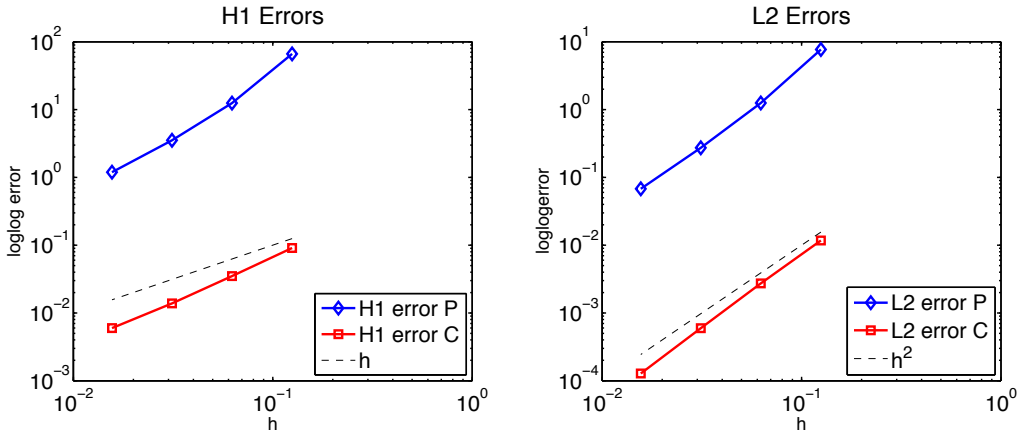


Figure 4.2 : H^1 versus h (left), L^2 versus h (right) for example 2.

Here we can see the DG methods in time and space are well suited for solving the miscible displacement equations. The convergence rates are as expected for both the pressure and concentration, in both the L^2 and H^1 norms. The code used here will be the used to solve for the pressure and concentration when solving the optimal control of the miscible displacement equations.

4.3.2 Two dimensions

Next consider simulations obtained from solving the miscible displace equations using the software DUNE. We consider the quarter five-spot problem, in which a fluid is injected from one corner, the bottom left $(0, 0)$ of a square domain to another corner, the top right $(1, 1)$, discussed in [14]. The injection rate varies from $0.16\, m/s^2$ to $0.18\, m/s^2$ and look at the concentration at two different time steps: $T = 0.5, 0.8$ seconds. We use the following parameters:

$$\phi = 0.2,$$

$$\alpha_l = 1.8 \times 10^{-5} m,$$

$$\alpha_t = 1.8 \times 10^{-6} m,$$

$$d_m = 1.8 \times 10^{-7} m/s^2.$$

The mobility ratio is defined to be 5.8/2.9 and we assume the gravity is negligible. The permeability matrix is assumed to be the identity matrix. We consider a problem with zero Dirichlet and Neumann boundary conditions for the concentration and zero Dirichlet and nonzero Neumann boundary conditions for the pressure. DG1 is used to discretize the PDEs in space for both the concentration and pressure and backward Euler is used to discretize the PDE in time. The examples are run on the DaVInCI cluster using one node, with a 16×16 grid, with no refinement and a time step size of $\Delta t = .1s$.

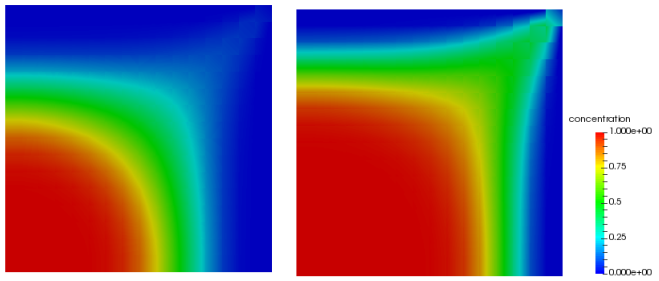


Figure 4.3 : Quarter five spot with injection rate $0.16\,m/s^2$ at $T = .5, .8$.

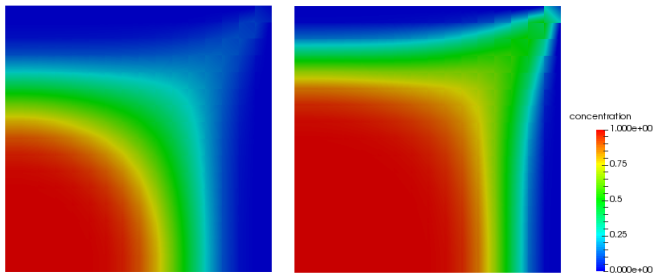


Figure 4.4 : Quarter five spot with injection rate $0.17\,m/s^2$ at $T = .5, .8$.

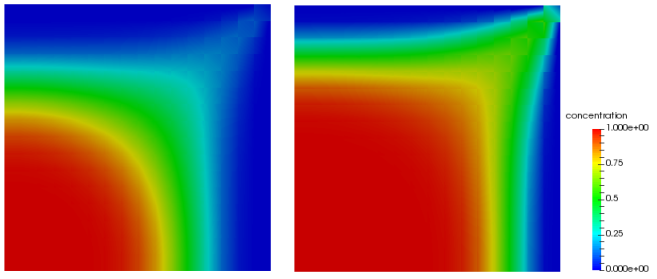


Figure 4.5 : Quarter five spot with injection rate $0.18\,m/s^2$ at $T = .5, .8$.

From Figures 4.3-4.5, we can see the affect of different injection rates on the quarter five-spot problem. As the injection rate increases, the faster the fluid moves from the injection site to the production site. We look at the concentration of the injected fluid at the $T = .5s$ and $T = .8s$ for each example over the line from the

bottom left corner $(0, 0)$ to the top right corner $(1, 1)$ of the domain. The line over which we are comparing the concentrations can be seen in Figure 4.6.

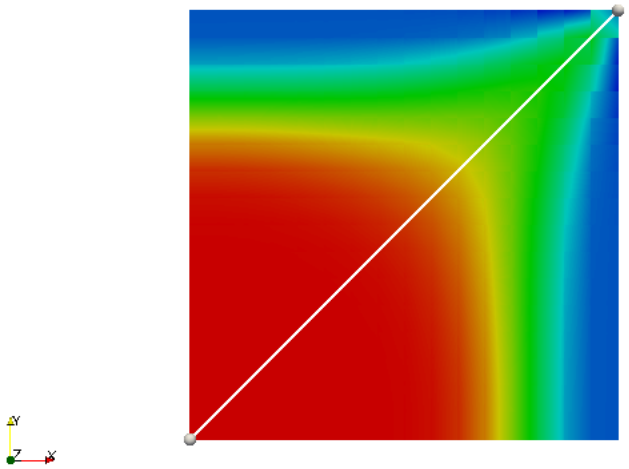


Figure 4.6 : Diagram of concentration cross section compared.

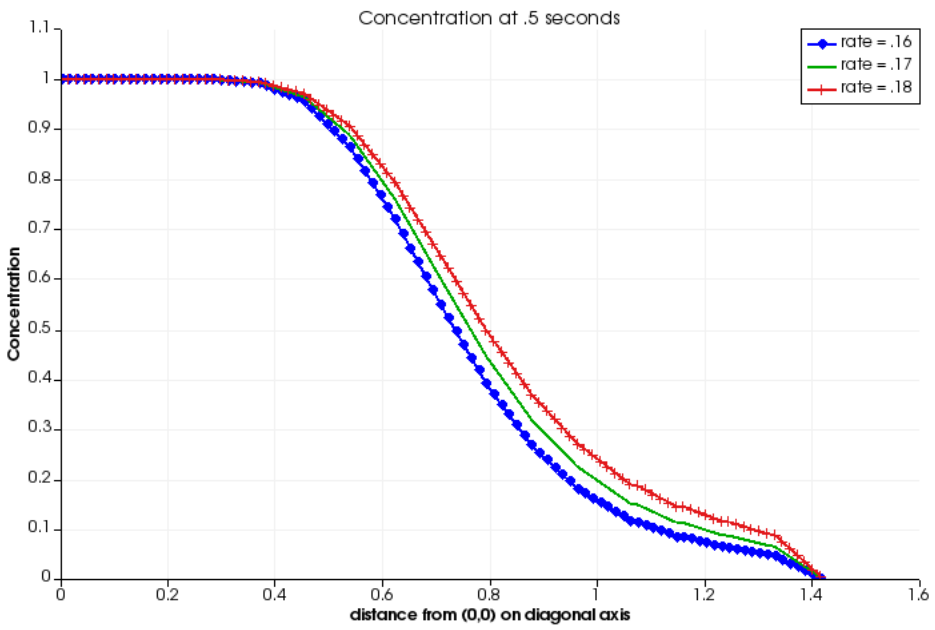


Figure 4.7 : Concentration from quarter five spot with varying injection rates at $T = .5s$

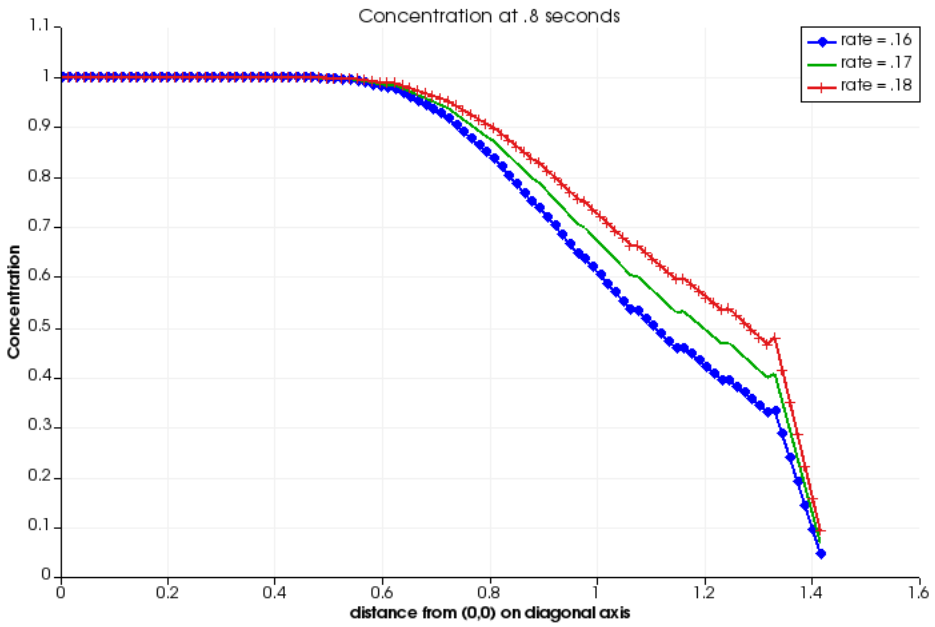


Figure 4.8 : Concentration from quarter five spot with varying injection rates at $T = .8s$

In Figure 4.7 and Figure 4.8, we can see that at $T = .5s$ and $T = .8s$, as the injection rate increases, the amount of space with concentration one increases. There is more of the injected fluid over the domain for the examples with higher injection rates.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this thesis, we have solved an optimal control problem with a system of steady-state PDEs using discontinuous Galerkin methods. We derived the optimality conditions for the optimal control problem. We used this spatial discretization to eliminate any instability with having a convection-dominated PDE, since continuous finite element methods can lead to oscillations when a small diffusion coefficient is present. We proved error estimates for the control with the discontinuous Galerkin discretization and we gave numerical results that validated the error estimates. It is clear that DG methods work well for solving optimal control problems governed by a system of linear, steady-state PDEs that are convection dominated. The convergence rates are optimal for both the states and the control.

We also solved an optimal control problem with a transport PDE. We derived the optimality conditions for this problem and discretized using a discontinuous Galerkin method in space and trapezoid method in time. We used a Newton Conjugate Gradient Method to solve the optimization problem. Last, we gave numerical results to validate the accuracy of the method. The convergence rates for the problem were suboptimal, but the numerical solution did converge to the exact solution for the state, control, and the adjoint state.

We gave the problem statement for the optimal control the miscible displacement equations, where the control is the flow rate of the injected fluid. We derived the optimality conditions for this problem. After stating a discontinuous Galerkin method

discretization of the PDEs, we gave numerical results to show that DG methods are well suited for solving the miscible displacement equations.

5.2 Future work

We will study the optimal control of the miscible displacement equations in much more detail. Though we have already derived the optimal conditions, we will focus on solving the optimization problem with the miscible displacement equations as PDE constraints numerically in one spatial dimension using a software in Matlab. We will then study the problem more in two and three dimensions using the software DUNE. We will prove derive error estimates for the control and states in this problem and run two and three dimensional simulations using DUNE.

Appendix A

A.1 Proofs from chapter 2

Before proving the lemmas from Chapter 2, we state and prove two lemmas needed first.

Lemma A.1

Let e be an face in Γ_h and let E_1 and E_2 be the elements connected by e. Then for all y in $V_h(\Omega)$, there exists a C independent of h such that

$$||[y - P_h y]||^2_{L^2(e)} \leq Ch^{2k+1} \left(|y|_{H^{k+1}(E_1)} + |y|_{H^{k+1}(E_2)} \right)^2 . \tag{A.1}$$

Proof A.1

$$\begin{aligned} ||[y - P_h y]||^2_{L^2(e)} &= ||(y - P_h y)|_{E_1} - (y - P_h y)|_{E_2}||^2_{L^2(e)} \\ &\leq \left(||(y - P_h y)|_{E_1}||_{L^2(e)} + ||(y - P_h y)|_{E_2}||_{L^2(e)} \right)^2 . \end{aligned} \tag{A.2}$$

From the trace lemma and (A.2), we obtain

$$\begin{aligned} \left(||(y - P_h y)|_{E_1}||_{L^2(e)} + ||(y - P_h y)|_{E_2}||_{L^2(e)} \right)^2 & \\ \leq \Bigg(\left(Ch^{-1}||y - P_h y||^2_{L^2(E_1)} + Ch||\nabla(y - P_h y)||^2_{L^2(E_1)} \right)^{1/2} & \\ + \left(Ch^{-1}||y - P_h y||^2_{L^2(E_1)} + Ch||\nabla(y - P_h y)||^2_{L^2(E_1)} \right)^{1/2} \Bigg)^2 & \\ \leq \Bigg(\left(Ch^{2k+1}|y|^2_{H^{k+1}(E_1)} \right)^{1/2} + \left(Ch^{2k+1}|y|^2_{H^{k+1}(E_2)} \right)^{1/2} \Bigg)^2 & \\ \leq \left(Ch^{k+1/2}|y|_{H^{k+1}(E_1)} + Ch^{k+1/2}|y|_{H^{k+1}(E_2)} \right)^2 & \\ \leq Ch^{2k+1} \left(|y|_{H^{k+1}(E_1)} + |y|_{H^{k+1}(E_2)} \right)^2 . & \end{aligned}$$

□

Lemma A.2

Given ϵ and $\mathbf{c}$ from the PDE, then for all y in $H^{k+1}(\Omega)$, there exists a C independent of h and ϵ such that

$$|||y - P_h y||| \leq Ch^k |y|_{H^{k+1}(\Omega)} \left(\epsilon^{1/2} + ||\mathbf{c}||_\infty^{1/2} h^{1/2}\right). \tag{A.3}$$

Proof A.2 From the definition of the DG norm, we have

$$\begin{aligned}
|||y - P_h y|||^2 &= \epsilon |||y - P_h y|||_{\text{diff}}^2 + |||y - P_h y|||_{\text{conv}}^2 \\
&= \epsilon \sum_{E \in E_h} ||\nabla(y - P_h y)||_{L^2(E)}^2 + \epsilon \sum_{e \in \Gamma_h \cup \partial\Omega} \frac{1}{h} ||[y - P_h y]||_{L^2(e)}^2 \\
&+ \frac{1}{2} \sum_{e \in \Gamma_h} \int_e |\mathbf{c} \cdot \mathbf{n}| [y - P_h y]^2 + \frac{1}{2} \sum_{e \in \partial\Omega_+} \int_e |\mathbf{c} \cdot \mathbf{n}| (y - P_h y)^2.
\end{aligned} \tag{A.4}$$

Recall the bounds for the L^2 errors for y in $H^{k+1}(E)$:

$$||y - P_h y||_{L^2(E)} \leq Ch^{k+1} |y|_{H^{k+1}(E)}, \tag{A.5}$$

$$||\nabla(y - P_h y)||_{L^2(E)} \leq Ch^k |y|_{H^{k+1}(E)}. \tag{A.6}$$

Combining (A.5) and (A.6) with (A.4), we obtain

$$\begin{aligned}
&\epsilon \sum_{E \in E_h} ||\nabla(y - P_h y)||_{L^2(E)}^2 + \epsilon \sum_{e \in \Gamma_h \cup \partial\Omega} \frac{1}{h} ||[y - P_h y]||_{L^2(e)}^2 + \frac{1}{2} \sum_{e \in \Gamma_h} \int_e |\mathbf{c} \cdot \mathbf{n}| [y - P_h y]^2 \\
&\qquad + \frac{1}{2} \sum_{e \in \partial\Omega_+} \int_e |\mathbf{c} \cdot \mathbf{n}| (y - P_h y)^2 \\
&\qquad \leq \epsilon \sum_{E \in E_h} Ch^{2k} |y|_{H^{k+1}(E)}^2 + \epsilon \sum_{e \in \Gamma_h \cup \partial\Omega} \frac{1}{h} ||[y - P_h y]||_{L^2(e)}^2 \\
&\qquad + \frac{1}{2} \sum_{e \in \Gamma_h} ||\mathbf{c}||_\infty ||[y - P_h y]||_{L^2(e)}^2 + \frac{1}{2} \sum_{e \in \partial\Omega_+} ||\mathbf{c}||_\infty C ||y - P_h y||_{L^2(e)}^2 \\
&\qquad \leq \epsilon \sum_{E \in E_h} Ch^{2k} |y|_{H^{k+1}(E)}^2 + \epsilon \sum_{e \in \Gamma_h \cup \partial\Omega} \frac{1}{h} ||[y - P_h y]||_{L^2(e)}^2 \\
&\qquad + \frac{1}{2} \sum_{e \in \Gamma_h} ||\mathbf{c}||_\infty ||[y - P_h y]||_{L^2(e)}^2 + \frac{1}{2} \sum_{e \in \partial\Omega_+} ||\mathbf{c}||_\infty Ch^{2k+1} |y|_{H^{k+1}(E)}^2.
\end{aligned}$$

From Lemma A.1, we can bound the other two terms:

$$\begin{aligned}
\epsilon \sum_{E\in E_h} Ch^{2k}|y|^2_{H^{k+1}(E)} + \epsilon \sum_{e\in\Gamma_h\cup\partial\Omega} \frac{1}{h}|||[y-P_hy]|||^2_{L^2(e)} + \frac{1}{2}\sum_{e\in\Gamma_h} ||\mathbf{c}||_\infty |||[y-P_hy]|||^2_{L^2(e)} \\
+ \frac{1}{2}\sum_{e\in\partial\Omega_+} ||\mathbf{c}||_\infty Ch^{2k+1}|y|^2_{H^{k+1}(E)} \\
\leq \epsilon \sum_{E\in E_h} Ch^{2k}|y|^2_{H^{k+1}(E)} + \epsilon \sum_{e\in\Gamma_h\cup\partial\Omega} \frac{1}{h}\left(Ch^{2k+1}\left(|y|_{H^{k+1}(E_1)} + |y|_{H^{k+1}(E_2)}\right)^2\right) \\
+ \frac{1}{2}\sum_{e\in\Gamma_h} ||\mathbf{c}||_\infty \left(Ch^{2k+1}\left(|y|_{H^{k+1}(E_1)} + |y|_{H^{k+1}(E_2)}\right)^2\right) \\
+ \frac{1}{2}\sum_{e\in\partial\Omega_+} ||\mathbf{c}||_\infty Ch^{2k+1}|y|^2_{H^{k+1}(E)}. \qquad \text{(A.7)}
\end{aligned}$$

We can then simplify (A.7) to obtain

$$\begin{aligned}
\epsilon \sum_{E\in E_h} Ch^{2k}|y|^2_{H^{k+1}(E)} + \epsilon \sum_{e\in\Gamma_h\cup\partial\Omega} \frac{1}{h}\left(Ch^{2k+1}\left(|y|_{H^{k+1}(E_1)} + |y|_{H^{k+1}(E_2)}\right)^2\right) \\
+ \frac{1}{2}\sum_{e\in\Gamma_h} ||\mathbf{c}||_\infty \left(Ch^{2k+1}\left(|y|_{H^{k+1}(E_1)} + |y|_{H^{k+1}(E_2)}\right)^2\right) \\
+ \frac{1}{2}\sum_{e\in\partial\Omega_+} ||\mathbf{c}||_\infty Ch^{2k+1}|y|^2_{H^{k+1}(E)} \\
\leq \epsilon Ch^{2k}|y|^2_{H^{k+1}(\Omega)} + \epsilon Ch^{2k}|y|^2_{H^{k+1}(\Omega)} + \frac{1}{2}||\mathbf{c}||_\infty Ch^{2k+1}|y|^2_{H^{k+1}(\Omega)} \\
+ \frac{1}{2}||\mathbf{c}||_\infty Ch^{2k+1}|y|^2_{H^{k+1}(\Omega)} \\
\leq Ch^{2k}|y|^2_{H^{k+1}(\Omega)}\left(\epsilon + ||\mathbf{c}||_\infty h\right).
\end{aligned}$$

□

Lemma A.3

There exists a C independent of h and ϵ such that

$$|||p_z - \tilde{p}_{z_h}|||^2_{\text{diff}} \leq Ch^{2k}|p_z|^2_{H^{k+1}(\Omega)} + C||z-\tilde{z}_h||^2_{L^2(\Omega)}, \qquad \text{(A.8)}$$

$$|||p_y - \tilde{p}_{y_h}|||^2 \leq Ch^{2k}|p_y|^2_{H^{k+1}(\Omega)}\left(\epsilon + ||\mathbf{c}||_\infty h + \epsilon^2 + \frac{||\mathbf{c}||^2_\infty}{\epsilon}\right) + \frac{C}{\epsilon}||\tilde{y}_h - y||^2_{L^2(\Omega)}. \qquad \text{(A.9)}$$

Proof A.3 Proof of (A.8)

From (2.41) and (2.36), we have

$$a_{\text{diff}}(v_h, \tilde{p}_{z_h}) = (\hat{z} - \tilde{z}_h, v_h), \tag{A.10}$$

$$a_{\text{diff}}(v_h, p_z) = (\hat{z} - z, v_h). \tag{A.11}$$

Subtracting (A.10) from (A.11), we have

$$a_{\text{diff}}(v_h, p_z - \tilde{p}_{z_h}) = (\hat{z} - z, v_h) - (\hat{z} - \tilde{z}_h, v_h) \tag{A.12}$$

$$= (\tilde{z}_h - z, v_h). \tag{A.13}$$

We want to bound $|||p_z - \tilde{p}_{z_h}|||_{\text{diff}}$. First, we use the triangle inequality.

$$|||p_z - \tilde{p}_{z_h}|||_{\text{diff}} \leq |||p_z - P_h p_z|||_{\text{diff}} + |||P_h p_z - \tilde{p}_{z_h}|||_{\text{diff}}. \tag{A.14}$$

From the error of the L^2 projection, we have a bound for $|||p_z - P_h p_z|||_{\text{diff}}$.

$$|||p_z - P_h p_z|||_{\text{diff}} \leq C h^k |p_z|_{H^{k+1}(\Omega)}. \tag{A.15}$$

Next, we will bound $|||P_h p_z - \tilde{p}_{z_h}|||_{\text{diff}}$. Let v_h be defined below,

$$v_h = P_h p_z - \tilde{p}_{z_h}.$$

Using the definition of v_h and (A.13), we add and subtract p_z to obtain

$$\begin{aligned} a_{\text{diff}}(P_h p_z - \tilde{p}_{z_h}, P_h p_z - \tilde{p}_{z_h}) &= a_{\text{diff}}(P_h p_z - p_z + p_z - \tilde{p}_{z_h}, P_h p_z - \tilde{p}_{z_h}) \\ &= a_{\text{diff}}(p_z - \tilde{p}_{z_h}, P_h p_z - \tilde{p}_{z_h}) + a_{\text{diff}}(P_h p_z - p_z, P_h p_z - \tilde{p}_{z_h}) \\ &= a_{\text{diff}}(P_h p_z - \tilde{p}_{z_h}, P_h p_z - p_z) + (\tilde{z}_h - z, P_h p_z - \tilde{p}_{z_h}). \end{aligned} \tag{A.16}$$

From coercivity of a_{diff}, we have

$$\begin{aligned} C|||P_h p_z - \tilde{p}_{z_h}|||^2_{\text{diff}} &\leq C h^k |p_z|_{H^{k+1}(\Omega)} |||P_h p_z - \tilde{p}_{z_h}|||_{\text{diff}} + ||\tilde{z}_h - z||_{L^2(\Omega)} ||P_h p_z - \tilde{p}_{z_h}||_{L^2(\Omega)} \\ &\leq C h^k |p_z|_{H^{k+1}(\Omega)} |||P_h p_z - \tilde{p}_{z_h}|||_{\text{diff}} + ||\tilde{z}_h - z||_{L^2(\Omega)} C |||P_h p_z - \tilde{p}_{z_h}|||_{\text{diff}}. \end{aligned} \tag{A.17}$$

We can simplify (A.17) to obtain

$$|||P_h p_z - \tilde{p}_{z_h}|||_{\text{diff}} \leq Ch^k |p_z|_{H^{k+1}(\Omega)} + ||\tilde{z}_h - z||_{L^2(\Omega)},$$
$$|||P_h p_z - \tilde{p}_{z_h}|||^2_{\text{diff}} \leq (Ch^k |p_z|_{H^{k+1}(\Omega)} + C||\tilde{z}_h - z||_{L^2(\Omega)})^2. \tag{A.18}$$

Combining (A.18) and (A.15) with (A.14), we obtain the error estimate.

$$\begin{aligned} |||p_z - \tilde{p}_{z_h}|||^2_{\text{diff}} &\leq Ch^{2k} |p_z|^2_{H^{k+1}(\Omega)} + C||\tilde{z}_h - z||^2_{L^2(\Omega)} + Ch^{2k} |p_z|^2_{H^{k+1}(\Omega)} \\ &\leq Ch^{2k} |p_z|^2_{H^{k+1}(\Omega)} + C||\tilde{z}_h - z||^2_{L^2(\Omega)}. \end{aligned}$$

Proof of (A.9):

We want to bound $|||p_y - \tilde{p}_{y_h}|||$. First we use the triangle inequality.

$$|||p_y - \tilde{p}_{y_h}||| \leq |||p_y - P_h p_y||| + |||P_h p_y - \tilde{p}_{y_h}|||.$$

From Lemma A.2, we have a bound for $|||p_y - P_h p_y|||$.

$$|||p_y - P_h p_y|||^2 \leq Ch^{2k} |p_y|^2_{H^{k+1}(\Omega)} \left(\epsilon + ||\mathbf{c}||_\infty h\right). \tag{A.19}$$

Next, we will bound $|||P_h p_y - \tilde{p}_{y_h}|||$:

From (2.40) and (2.35), we have

$$\epsilon a_{\text{diff}}(v_h, \tilde{p}_{y_h}) + a_{\text{conv}}(v_h, \tilde{p}_{y_h}) = (\hat{y} - \tilde{y}_h, v_h), \tag{A.20}$$
$$\epsilon a_{\text{diff}}(v_h, p_y) + a_{\text{conv}}(v_h, p_y) = (\hat{y} - y_h, v_h). \tag{A.21}$$

Subtracting (A.21) from (A.20), we have

$$\begin{aligned} \epsilon a_{\text{diff}}(v_h, \tilde{p}_{y_h} - P_h p_y) + a_{\text{conv}}(v_h, \tilde{p}_{y_h} - P_h p_y) &= \epsilon a_{\text{diff}}(v_h, p_y - P_h p_y) \\ &\quad + a_{\text{conv}}(v_h, p_y - P_h p_y) + (y_h - \tilde{y}_h, v_h). \end{aligned} \tag{A.22}$$

Next, we let v_h be defined as follows

$$v_h = \tilde{p}_{y_h} - P_h p_y.$$

From coercivity of a_{diff}, we have

$$C|||\tilde{p}_{y_h} - P_h p_y|||^2 \leq \epsilon a_{\text{diff}}(\tilde{p}_{y_h} - P_h p_y, \tilde{p}_{y_h} - P_h p_y) + a_{\text{conv}}(\tilde{p}_{y_h} - P_h p_y, \tilde{p}_{y_h} - P_h p_y).$$

From the definition of v_h and (A.22), we have

$$\begin{aligned}
\epsilon a_{\text{diff}}(\tilde{p}_{y_h} - P_h p_y, \tilde{p}_{y_h} - P_h p_y) + a_{\text{conv}}(\tilde{p}_{y_h} - P_h p_y, \tilde{p}_{y_h} - P_h p_y) & \\
&= \epsilon a_{\text{diff}}(\tilde{p}_{y_h} - P_h p_y, p_y - P_h p_y) + a_{\text{conv}}(\tilde{p}_{y_h} - P_h p_y, p_y - P_h p_y) \\
&\quad + (y_h - \tilde{y}_h, \tilde{p}_{y_h} - P_h p_y) \\
&\leq \epsilon a_{\text{diff}}(\tilde{p}_{y_h} - P_h p_y, p_y - P_h p_y) + a_{\text{conv}}(\tilde{p}_{y_h} - P_h p_y, p_y - P_h p_y) \\
&\quad + ||y - \tilde{y}_h||_{L^2(\Omega)} ||\tilde{p}_{y_h} - P_h p_y||_{L^2(\Omega)} \qquad \text{(A.23)}
\end{aligned}$$

We can bound the a_{diff} term.

$$\epsilon a_{\text{diff}}(\tilde{p}_{y_h} - P_h p_y, p_y - P_h p_y) \leq \epsilon C h^k |p_y|_{H^{k+1}(\Omega)} |||\tilde{p}_{y_h} - P_h p_y|||_{\text{diff}}. \qquad \text{(A.24)}$$

We can also bound the a_{conv} term using the definition.

$$\begin{aligned}
a_{\text{conv}}(\tilde{p}_{y_h} - P_h p_y, p_y - P_h p_y) &= -\sum_{E \in E_h} \int_E (\tilde{p}_{y_h} - P_h p_y)(\nabla(p_y - P_h p_y) \cdot \mathbf{c}) \\
&\quad + \sum_{e \in \Gamma_h} \int_e (\tilde{p}_{y_h} - P_h p_y)^{\text{up}} (\mathbf{c} \cdot \mathbf{n}_e)[p_y - P_h p_y] \\
&\quad + \int_{\partial\Omega_+} (\tilde{p}_{y_h} - P_h p_y)(p_y - P_h p_y)(\mathbf{c} \cdot \mathbf{n}_e) \\
&\leq C||\mathbf{c}||_\infty ||\tilde{p}_{y_h} - P_h p_y||_{L^2(\Omega)} \left(\left(\sum_{E \in E_h} ||\nabla(p_y - P_h p_y)||^2_{L^2(E)} \right)^{1/2} \right. \\
&\quad \left. + \left(\sum_{e \in \Gamma_h \cup \partial\Omega_+} \frac{1}{h} |||[p_y - P_h p_y]]|||^2_{L^2(e)} \right)^{1/2} \right) \\
&\leq C||\mathbf{c}||_\infty \frac{1}{\sqrt{\epsilon}} |||\tilde{p}_{y_h} - P_h p_y||| h^k |p_y|_{H^{k+1}(\Omega)} \qquad \text{(A.25)}
\end{aligned}$$

We combine (A.23), (A.24) and (A.25):

$$
\begin{aligned}
|||\tilde{p}_{y_h} - P_h p_y|||^2 &\leq \epsilon C h^k |p_y|_{H^{k+1}(\Omega)} |||\tilde{p}_{y_h} - P_h p_y||| + \frac{C}{\sqrt{\epsilon}} ||\mathbf{c}||_\infty |||\tilde{p}_{y_h} - P_h p_y||| h^k |p_y|_{H^{k+1}(\Omega)} \\
&\quad + ||y - \tilde{y}_h||_{L^2(\Omega)} ||\tilde{p}_{y_h} - P_h p_y||_{L^2(\Omega)} \\
&\leq \epsilon C h^k |p_y|_{H^{k+1}(\Omega)} |||\tilde{p}_{y_h} - P_h p_y||| + \frac{C}{\sqrt{\epsilon}} ||\mathbf{c}||_\infty |||\tilde{p}_{y_h} - P_h p_y||| h^k |p_y|_{H^{k+1}(\Omega)} \\
&\quad + \frac{C}{\sqrt{\epsilon}} ||y - \tilde{y}_h||_{L^2(\Omega)} |||\tilde{p}_{y_h} - P_h p_y|||, \qquad \text{(A.26)}
\end{aligned}
$$

We simplify (A.26) to obtain

$$
|||\tilde{p}_{y_h} - P_h p_y||| \leq C h^k |p_y|_{H^{k+1}(\Omega)} \left(\epsilon + \frac{||\mathbf{c}||_\infty}{\sqrt{\epsilon}} \right) + \frac{C}{\sqrt{\epsilon}} ||y - \tilde{y}_h||_{L^2(\Omega)}. \qquad \text{(A.27)}
$$

We obtain the bound on $|||\tilde{p}_{y_h} - P_h p_y|||$:

$$
|||\tilde{p}_{y_h} - P_h p_y|||^2 \leq C h^{2k} |p_y|^2_{H^{k+1}(\Omega)} \left(\epsilon^2 + \frac{||\mathbf{c}||^2_\infty}{\epsilon} \right) + \frac{C}{\epsilon} ||y - \tilde{y}_h||^2_{L^2(\Omega)}. \qquad \text{(A.28)}
$$

Combining (A.19) and (A.28), we obtain the desired result.

$$
\begin{aligned}
|||p_y - \tilde{p}_{y_h}|||^2 &\leq C |||p_y - P_h p_y|||^2 + C |||\tilde{p}_{y_h} - P_h p_y|||^2 \\
&\leq C h^{2k} |p_y|^2_{H^{k+1}(\Omega)} (\epsilon + ||\mathbf{c}||_\infty h) + C h^{2k} |p_y|^2_{H^{k+1}(\Omega)} \left(\epsilon^2 + \frac{||\mathbf{c}||^2_\infty}{\epsilon} \right) \\
&\quad + \frac{C}{\epsilon} ||y - \tilde{y}_h||^2_{L^2(\Omega)} \\
&\leq C h^{2k} |p_y|^2_{H^{k+1}(\Omega)} \left(\epsilon + ||\mathbf{c}||_\infty h + \epsilon^2 + \frac{||\mathbf{c}||^2_\infty}{\epsilon} \right) + \frac{C}{\epsilon} ||\tilde{y}_h - y||^2_{L^2(\Omega)}.
\end{aligned}
$$

□

Lemma A.4

There exists a C independent of h and ϵ such that

$$
|||\tilde{y}_h - P_h y||| \leq C h^k |y|_{H^{k+1}(\Omega)}, \qquad \text{(A.29)}
$$

$$
|||\tilde{z}_h - P_h z|||_{\text{diff}} \leq C h^k |z|_{H^{k+1}(\Omega)}. \qquad \text{(A.30)}
$$

Proof A.4 Proof of (A.30):

From (2.39) and (2.33), we have

$$a_{\text{diff}}(\tilde{z}_h, v_h) = (u, v_h) + l_g(v_h), \tag{A.31}$$

$$a_{\text{diff}}(z, v_h) = (u, v_h) + l_g(v_h). \tag{A.32}$$

We subtract (A.32) from (A.31) and let $v_h = \tilde{z}_h - P_h z$.

$$a_{\text{diff}}(\tilde{z}_h - z, \tilde{z}_h - P_h z) = 0. \tag{A.33}$$

We add and subtract z to obtain

$$\begin{aligned} a_{\text{diff}}(\tilde{z}_h - P_h z, \tilde{z}_h - P_h z) &= a_{\text{diff}}(\tilde{z}_h - z + z - P_h z, \tilde{z}_h - P_h z) \\ &= a_{\text{diff}}(z - P_h z, \tilde{z}_h - P_h z) + a_{\text{diff}}(\tilde{z}_h - z, \tilde{z}_h - P_h z) \\ &= a_{\text{diff}}(z - P_h z, \tilde{z}_h - P_h z) \\ &\leq C|||\tilde{z}_h - P_h z|||_{\text{diff}} h^k |z|_{H^{k+1}(\Omega)}. \end{aligned} \tag{A.34}$$

This gives us the bound on $|||\tilde{z}_h - P_h z|||_{\text{diff}}$:

$$|||\tilde{z}_h - P_h z|||_{\text{diff}} \leq C h^k |z|_{H^{k+1}(\Omega)}. \tag{A.35}$$

Proof of (A.29):

From (2.38) and (2.32), we have

$$\epsilon a_{\text{diff}}(\tilde{y}_h, v_h) + a_{\text{conv}}(\tilde{y}_h, v_h) = (u, v_h) + l_f(v_h), \tag{A.36}$$

$$\epsilon a_{\text{diff}}(y, v_h) + a_{\text{conv}}(y, v_h) = (u, v_h) + l_f(v_h). \tag{A.37}$$

We subtract (A.37) from (A.36) to obtain

$$\epsilon a_{\text{diff}}(\tilde{y}_h - y, v_h) + a_{\text{conv}}(\tilde{y}_h - y, v_h) = 0. \tag{A.38}$$

We add and subtract y in a_{diff} and a_{conv}:

$$\begin{aligned}\epsilon a_{\text{diff}}(\tilde{y}_h - P_h y, v_h) + a_{\text{conv}}(\tilde{y}_h - P_h y, v_h) &= \epsilon a_{\text{diff}}(\tilde{y}_h - y + y - P_h y, v_h) \\ &\quad + a_{\text{conv}}(\tilde{y}_h - y + y - P_h y, v_h) \\ &= \epsilon a_{\text{diff}}(y - P_h y, v_h) + a_{\text{conv}}(y - P_h y, v_h) \\ &\quad + \epsilon a_{\text{diff}}(\tilde{y}_h - y, v_h) + a_{\text{conv}}(\tilde{y}_h - y, v_h) \\ &= \epsilon a_{\text{diff}}(y - P_h y, v_h) + a_{\text{conv}}(y - P_h y, v_h).\end{aligned} \tag{A.39}$$

We let $v_h = \tilde{y}_h - P_h y$ to obtain

$$\begin{aligned}\epsilon a_{\text{diff}}(\tilde{y}_h - P_h y, \tilde{y}_h - P_h y) + a_{\text{conv}}(\tilde{y}_h - P_h y, \tilde{y}_h - P_h y) &= \epsilon a_{\text{diff}}(y - P_h y, \tilde{y}_h - P_h y) \\ &\quad + a_{\text{conv}}(y - P_h y, \tilde{y}_h - P_h y).\end{aligned} \tag{A.40}$$

We use the coercivity of a_{diff} and (A.40) to obtain

$$\begin{aligned}C\epsilon|||\tilde{y}_h - P_h y|||^2_{\text{diff}} + |||\tilde{y}_h - P_h y|||^2_{\text{conv}} &\leq \epsilon a_{\text{diff}}(y - P_h y, \tilde{y}_h - P_h y) + a_{\text{conv}}(y - P_h y, \tilde{y}_h - P_h y) \\ &\leq \epsilon C|||\tilde{y}_h - P_h y|||_{\text{diff}} h^k |y|_{H^{k+1}(\Omega)} \\ &\quad + a_{\text{conv}}(y - P_h y, \tilde{y}_h - P_h y).\end{aligned} \tag{A.41}$$

We use the definition of a_{conv} and Cauchy-Schwarz's inequality to bound (A.41), which can be seen from (A.25):

$$a_{\text{conv}}(\tilde{p}_{y_h} - P_h p_y, p_y - P_h p_y) \leq \frac{C}{\sqrt{\epsilon}} h^k ||\mathbf{c}||_\infty |||\tilde{p}_{y_h} - P_h p_y|||| p_y|_{H^{k+1}(\Omega)}. \tag{A.42}$$

This gives us a bound on $a_{\text{conv}}(y - P_h y, \tilde{y}_h - P_h y)$:

$$a_{\text{conv}}(y - P_h y, \tilde{y}_h - P_h y) \leq \frac{C}{\sqrt{\epsilon}} h^k ||\mathbf{c}||_\infty |||\tilde{y}_h - P_h y|||| y|_{H^{k+1}(\Omega)}. \tag{A.43}$$

We combine (A.43) with (A.41) to obtain

$$\begin{aligned} C\epsilon|||\tilde{y}_h - P_h y|||^2_{\text{diff}} + |||\tilde{y}_h - P_h y|||^2_{\text{conv}} &\leq \epsilon C h^k |||\tilde{y}_h - P_h y|||_{\text{diff}} |y|_{H^{k+1}(\Omega)} \\ &\quad + \frac{C}{\sqrt{\epsilon}} h^k ||\mathbf{c}||_\infty |||\tilde{y}_h - P_h y||| \, |y|_{H^{k+1}(\Omega)} \\ &\leq C h^k |y|_{H^{k+1}(\Omega)} \left(\epsilon |||\tilde{y}_h - P_h y|||_{\text{diff}} \right. \\ &\quad \left. + \frac{||\mathbf{c}||_\infty}{\sqrt{\epsilon}} |||\tilde{y}_h - P_h y||| \right) \\ &\leq C h^k |y|_{H^{k+1}(\Omega)} |||\tilde{y}_h - P_h y||| \left(\sqrt{\epsilon} + \frac{||\mathbf{c}||_\infty}{\sqrt{\epsilon}} \right). \end{aligned} \tag{A.44}$$

From (A.44), we have the bound on $|||\tilde{y}_h - P_h y|||$:

$$|||\tilde{y}_h - P_h y||| \leq C h^k \left(\sqrt{\epsilon} + \frac{||\mathbf{c}||_\infty}{\sqrt{\epsilon}} \right) |y|_{H^{k+1}(\Omega)}.$$

□

A.2 Proof from chapter 4

Recall the definition of $D(\mathbf{u})$ and $E(\mathbf{u})$:

$$D(\mathbf{u}) = d_m I + |\mathbf{u}| \left(\alpha_l E(\mathbf{u}) + \alpha_t (I - E(\mathbf{u})) \right), \tag{A.45}$$

$$E(\mathbf{u}) = \mathbf{u}\mathbf{u}^T |\mathbf{u}|^{-2}. \tag{A.46}$$

Define the function $f : \mathbb{R}^d \to \mathbb{R}$ by:

$$f(\mathbf{u}) = |\mathbf{u}|. \tag{A.47}$$

To determine the optimality conditions for the miscible displacement problem, $D'(\mathbf{u})(\bar{\mathbf{u}})$ needs to be computed. First, we prove a few results we will need for the derivative of $D(\mathbf{u})$.

Proposition A.1

The derivative of f in the direction of $\bar{\mathbf{u}}$ for $\mathbf{u} \neq 0$ is defined as:

$$f'(\mathbf{u})(\bar{\mathbf{u}}) = \frac{\mathbf{u}^T \bar{\mathbf{u}}}{|\mathbf{u}|}.$$

Proof A.5

$$
\begin{aligned}
f'(\mathbf{u})(\bar{\mathbf{u}}) &= \lim_{t\to 0} \frac{f(\mathbf{u}+t\bar{\mathbf{u}}) - f(\mathbf{u})}{t} \\
&= \lim_{t\to 0} \frac{|\mathbf{u}+t\bar{\mathbf{u}}| - |\mathbf{u}|}{t} \\
&= \lim_{t\to 0} \frac{\sqrt{(\mathbf{u}+t\bar{\mathbf{u}})^T(\mathbf{u}+t\bar{\mathbf{u}})} - \sqrt{\mathbf{u}^T\mathbf{u}}}{t} \\
&= \lim_{t\to 0} \frac{\sqrt{(\mathbf{u}+t\bar{\mathbf{u}})^T(\mathbf{u}+t\bar{\mathbf{u}})} - \sqrt{\mathbf{u}^T\mathbf{u}}}{t} \left(\frac{\sqrt{(\mathbf{u}+t\bar{\mathbf{u}})^T(\mathbf{u}+t\bar{\mathbf{u}})} + \sqrt{\mathbf{u}^T\mathbf{u}}}{\sqrt{(\mathbf{u}+t\bar{\mathbf{u}})^T(\mathbf{u}+t\bar{\mathbf{u}})} + \sqrt{\mathbf{u}^T\mathbf{u}}} \right) \\
&= \lim_{t\to 0} \frac{(\mathbf{u}+t\bar{\mathbf{u}})^T(\mathbf{u}+t\bar{\mathbf{u}}) - \mathbf{u}^T\mathbf{u}}{t\left(\sqrt{(\mathbf{u}+t\bar{\mathbf{u}})^T(\mathbf{u}+t\bar{\mathbf{u}})} + \sqrt{\mathbf{u}^T\mathbf{u}} \right)} \\
&= \lim_{t\to 0} \frac{2t\bar{\mathbf{u}}^T\mathbf{u} + t^2\bar{\mathbf{u}}^T\bar{\mathbf{u}}}{t\left(|\mathbf{u}+t\bar{\mathbf{u}}| + |\mathbf{u}^T\mathbf{u}|\right)} \\
&= \lim_{t\to 0} \frac{2\bar{\mathbf{u}}^T\mathbf{u} + t\bar{\mathbf{u}}^T\bar{\mathbf{u}}}{|\mathbf{u}+t\bar{\mathbf{u}}| + |\mathbf{u}^T\mathbf{u}|} \\
&= \frac{\mathbf{u}^T\bar{\mathbf{u}}}{|\mathbf{u}|}.
\end{aligned}
$$

□

Proposition A.2

The derivative of $E(\mathbf{u})$ *in the direction of* $\bar{\mathbf{u}}$ *for* $\mathbf{u} \neq 0$ *is defined by:*

$$
E'(\mathbf{u})(\bar{\mathbf{u}}) = \frac{1}{|\mathbf{u}|^2}\left(\bar{\mathbf{u}}\mathbf{u}^T + \mathbf{u}\bar{\mathbf{u}}^T - 2\mathbf{u}^T\bar{\mathbf{u}}E(\mathbf{u})\right).
$$

Proof A.6

$$
\begin{aligned}
E'(\mathbf{u})(\bar{\mathbf{u}}) &= \lim_{t\to 0} \frac{E(\mathbf{u}+t\bar{\mathbf{u}}) - E(\mathbf{u})}{t} \\
&= \lim_{t\to 0} \frac{(\mathbf{u}+t\bar{\mathbf{u}})(\mathbf{u}+t\bar{\mathbf{u}})^T|\mathbf{u}+t\bar{\mathbf{u}}|^{-2} - \mathbf{u}\mathbf{u}^T|\mathbf{u}|^{-2}}{t} \\
&= \lim_{t\to 0} \frac{1}{t}\left(\frac{(\mathbf{u}+t\bar{\mathbf{u}})(\mathbf{u}+t\bar{\mathbf{u}})^T}{(\mathbf{u}+t\bar{\mathbf{u}})^T(\mathbf{u}+t\bar{\mathbf{u}})} - \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}}\right) \\
&= \lim_{t\to 0} \left(\frac{(\mathbf{u}+t\bar{\mathbf{u}})(\mathbf{u}+t\bar{\mathbf{u}})^T\left(\mathbf{u}^T\mathbf{u}\right) - \mathbf{u}\mathbf{u}^T(\mathbf{u}+t\bar{\mathbf{u}})^T(\mathbf{u}+t\bar{\mathbf{u}})}{t(\mathbf{u}+t\bar{\mathbf{u}})^T(\mathbf{u}+t\bar{\mathbf{u}})\left(\mathbf{u}^T\mathbf{u}\right)}\right) \\
&= \lim_{t\to 0} \left(\frac{\left(\mathbf{u}\mathbf{u}^T + t\mathbf{u}\bar{\mathbf{u}}^T + t\bar{\mathbf{u}}\mathbf{u}^T + t^2\bar{\mathbf{u}}\bar{\mathbf{u}}^T\right)\left(\mathbf{u}^T\mathbf{u}\right) - \mathbf{u}\mathbf{u}^T(\mathbf{u}^T\mathbf{u} + 2t\bar{\mathbf{u}}^T\mathbf{u} + t^2\bar{\mathbf{u}}^T\bar{\mathbf{u}})}{t(\mathbf{u}+t\bar{\mathbf{u}})^T(\mathbf{u}+t\bar{\mathbf{u}})\left(\mathbf{u}^T\mathbf{u}\right)}\right) \\
&= \lim_{t\to 0} \left(\frac{\left(t\mathbf{u}\bar{\mathbf{u}}^T + t\bar{\mathbf{u}}\mathbf{u}^T + t^2\bar{\mathbf{u}}\bar{\mathbf{u}}^T\right)\left(\mathbf{u}^T\mathbf{u}\right) - \mathbf{u}\mathbf{u}^T(2t\bar{\mathbf{u}}^T\mathbf{u} + t^2\bar{\mathbf{u}}^T\bar{\mathbf{u}})}{t(\mathbf{u}+t\bar{\mathbf{u}})^T(\mathbf{u}+t\bar{\mathbf{u}})\left(\mathbf{u}^T\mathbf{u}\right)}\right) \\
&= \lim_{t\to 0} \left(\frac{\left(\mathbf{u}\bar{\mathbf{u}}^T + \bar{\mathbf{u}}\mathbf{u}^T + t\bar{\mathbf{u}}\bar{\mathbf{u}}^T\right)\left(\mathbf{u}^T\mathbf{u}\right) - \mathbf{u}\mathbf{u}^T(2\bar{\mathbf{u}}^T\mathbf{u} + t\bar{\mathbf{u}}^T\bar{\mathbf{u}})}{(\mathbf{u}+t\bar{\mathbf{u}})^T(\mathbf{u}+t\bar{\mathbf{u}})\left(\mathbf{u}^T\mathbf{u}\right)}\right) \\
&= \frac{\left(\mathbf{u}\bar{\mathbf{u}}^T + \bar{\mathbf{u}}\mathbf{u}^T\right)\left(\mathbf{u}^T\mathbf{u}\right) - 2\mathbf{u}\mathbf{u}^T\bar{\mathbf{u}}^T\mathbf{u}}{\left(\mathbf{u}^T\mathbf{u}\right)\left(\mathbf{u}^T\mathbf{u}\right)} \\
&= \frac{\mathbf{u}\bar{\mathbf{u}}^T + \bar{\mathbf{u}}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} - 2\frac{\mathbf{u}\mathbf{u}^T\bar{\mathbf{u}}^T\mathbf{u}}{\left(\mathbf{u}^T\mathbf{u}\right)\left(\mathbf{u}^T\mathbf{u}\right)} \\
&= \frac{\mathbf{u}\bar{\mathbf{u}}^T + \bar{\mathbf{u}}\mathbf{u}^T}{|\mathbf{u}|^2} - 2E(\mathbf{u})\frac{\bar{\mathbf{u}}^T\mathbf{u}}{|\mathbf{u}|^2} \\
&= \frac{1}{|\mathbf{u}|^2}\left(\bar{\mathbf{u}}\mathbf{u}^T + \mathbf{u}\bar{\mathbf{u}}^T - 2\mathbf{u}^T\bar{\mathbf{u}}E(\mathbf{u})\right).
\end{aligned}
$$

□

Lemma A.5

The derivative of $D(\mathbf{u})$ in the direction of $\bar{\mathbf{u}}$ for $\mathbf{u} \neq 0$ is defined by:

$$
D'(\mathbf{u})(\bar{\mathbf{u}}) = \frac{\alpha_l - \alpha_t}{|\mathbf{u}|}\left(\bar{\mathbf{u}}\mathbf{u}^T + \mathbf{u}\bar{\mathbf{u}}^T\right) + \mathbf{u}^T\bar{\mathbf{u}}\left(\frac{(\alpha_t - \alpha_l)E(\mathbf{u}) + \alpha_t I}{|\mathbf{u}|}\right).
$$

Proof A.7 Let D_1 and D_2 be defined by:

$$
D_1(\mathbf{u}) = (\alpha_l - \alpha_t)|\mathbf{u}|E(\mathbf{u}),
$$

$$
D_2(\mathbf{u}) = \alpha_t|\mathbf{u}|I.
$$

Then we can write D as a sum of D_1 and D_2:

$$D(\mathbf{u}) = d_m I + D_1(\mathbf{u}) + D_2(\mathbf{u}). \tag{A.48}$$

Next we compute the derivative of D_1.

$$\begin{aligned}
D_1'(\mathbf{u})(\bar{\mathbf{u}}) &= \lim_{t\to 0} \frac{D_1(\mathbf{u}+t\bar{\mathbf{u}}) - D_1(\mathbf{u})}{t} \\
&= \lim_{t\to 0} \frac{(\alpha_l - \alpha_t)}{t} \left(|\mathbf{u}+t\bar{\mathbf{u}}|E(\mathbf{u}+\mathbf{t}\bar{\mathbf{u}}) - |\mathbf{u}|E(\mathbf{u})\right) \\
&= \lim_{t\to 0} \frac{(\alpha_l - \alpha_t)}{t} \left(|\mathbf{u}+t\bar{\mathbf{u}}|E(\mathbf{u}+\mathbf{t}\bar{\mathbf{u}}) \pm |\mathbf{u}+t\bar{\mathbf{u}}|E(\mathbf{u}) - |\mathbf{u}|E(\mathbf{u})\right) \\
&= \lim_{t\to 0} (\alpha_l - \alpha_t) \left(\frac{|\mathbf{u}+t\bar{\mathbf{u}}|(E(\mathbf{u}+\mathbf{t}\bar{\mathbf{u}}) - E(\mathbf{u}))}{t} + \frac{(|\mathbf{u}+t\bar{\mathbf{u}}| - |\mathbf{u}|)E(\mathbf{u})}{t}\right) \\
&= (\alpha_l - \alpha_t) \left(|\mathbf{u}|E'(\mathbf{u})(\bar{\mathbf{u}}) + f'(\mathbf{u})(\bar{\mathbf{u}})E(\mathbf{u})\right).
\end{aligned} \tag{A.49}$$

Next we compute the derivative of D_2.

$$\begin{aligned}
D_2'(\mathbf{u})(\bar{\mathbf{u}}) &= \lim_{t\to 0} \frac{D_2(\mathbf{u}+t\bar{\mathbf{u}}) - D_2(\mathbf{u})}{t} \\
&= \lim_{t\to 0} \frac{\alpha_t|\mathbf{u}+t\bar{\mathbf{u}}|I - \alpha_t|\mathbf{u}|I}{t} \\
&= \lim_{t\to 0} \alpha_t \frac{|\mathbf{u}+t\bar{\mathbf{u}}| - |\mathbf{u}|}{t} I \\
&= \alpha_t f'(\mathbf{u})(\bar{\mathbf{u}})I.
\end{aligned} \tag{A.50}$$

We now take the derivative of D in the direction of $\bar{\mathbf{u}}$ using (A.48).

$$D'(\mathbf{u})(\bar{\mathbf{u}}) = D_1'(\mathbf{u}) + D_2'(\mathbf{u}).$$

From (A.49) and (A.50), we can rewrite $D'(\mathbf{u})(\bar{\mathbf{u}})$:

$$D'(\mathbf{u})(\bar{\mathbf{u}}) = (\alpha_l - \alpha_t) \left(|\mathbf{u}|E'(\mathbf{u})(\bar{\mathbf{u}}) + f'(\mathbf{u})(\bar{\mathbf{u}})E(\mathbf{u})\right) + \alpha_t f'(\mathbf{u})(\bar{\mathbf{u}})I.$$

We now use Propositions A.1 and A.2 to compute $D'(\mathbf{u})(\bar{\mathbf{u}})$.

$$\begin{aligned}
D'(\mathbf{u})(\bar{\mathbf{u}}) &= (\alpha_l - \alpha_t)\left(|\mathbf{u}|\frac{1}{|\mathbf{u}|^2}\left(\bar{\mathbf{u}}\mathbf{u}^T + \mathbf{u}\bar{\mathbf{u}}^T - 2\mathbf{u}^T\bar{\mathbf{u}}E(\mathbf{u})\right) + \frac{\mathbf{u}^T\bar{\mathbf{u}}}{|\mathbf{u}|}E(\mathbf{u})\right) + \alpha_t\frac{\mathbf{u}^T\bar{\mathbf{u}}}{|\mathbf{u}|}I \\
&= (\alpha_l - \alpha_t)\left(\frac{1}{|\mathbf{u}|}\left(\bar{\mathbf{u}}\mathbf{u}^T + \mathbf{u}\bar{\mathbf{u}}^T\right) - \frac{\mathbf{u}^T\bar{\mathbf{u}}E(\mathbf{u})}{|\mathbf{u}|}\right) + \alpha_t\frac{\mathbf{u}^T\bar{\mathbf{u}}}{|\mathbf{u}|}I \\
&= \frac{\alpha_l - \alpha_t}{|\mathbf{u}|}\left(\bar{\mathbf{u}}\mathbf{u}^T + \mathbf{u}\bar{\mathbf{u}}^T\right) - (\alpha_l - \alpha_t)\frac{\mathbf{u}^T\bar{\mathbf{u}}E(\mathbf{u})}{|\mathbf{u}|} + \alpha_t\frac{\mathbf{u}^T\bar{\mathbf{u}}}{|\mathbf{u}|}I \\
&= \frac{\alpha_l - \alpha_t}{|\mathbf{u}|}\left(\bar{\mathbf{u}}\mathbf{u}^T + \mathbf{u}\bar{\mathbf{u}}^T\right) + \mathbf{u}^T\bar{\mathbf{u}}\left(\frac{(\alpha_t - \alpha_l)E(\mathbf{u}) + \alpha_t I}{|\mathbf{u}|}\right).
\end{aligned}$$

□

To determine the optimality condition when we set the derivative of the Lagrangian with respect to $\mathbf{u}$ equal to zero, we want to isolate $\bar{\mathbf{u}}$ in $(D'(\mathbf{u})(\bar{\mathbf{u}})\nabla c)\cdot\nabla\lambda_c$.

Lemma A.6

$$(D'(\mathbf{u})(\bar{\mathbf{u}})\nabla c)\cdot\nabla\lambda_c = \bar{\mathbf{u}}\cdot F(\mathbf{u},\nabla c,\nabla\lambda_c),$$

where we define F by

$$F(\mathbf{u},\nabla c,\nabla\lambda_c) = \frac{\alpha_l - \alpha_t}{|\mathbf{u}|}\left(\nabla\lambda_c\nabla c^T\mathbf{u} + \nabla c\nabla\lambda_c^T\mathbf{u}\right) + \mathbf{u}\nabla\lambda_c^T\left(\frac{(\alpha_t - \alpha_l)E(\mathbf{u}) + \alpha_t I}{|\mathbf{u}|}\right)\nabla c.$$

Proof A.8 We define the following parameters $b(\mathbf{u})$ and $B(\mathbf{u})$, which are independent of $\bar{\mathbf{u}}$:

$$b(\mathbf{u}) = \frac{\alpha_l - \alpha_t}{|\mathbf{u}|}, \tag{A.51}$$

$$B(\mathbf{u}) = \left(\frac{(\alpha_t - \alpha_l)E(\mathbf{u}) + \alpha_t I}{|\mathbf{u}|}\right). \tag{A.52}$$

We can rewrite the definition of $D'(\mathbf{u})(\bar{\mathbf{u}})$ using (A.51) and (A.52):

$$D'(\mathbf{u})(\bar{\mathbf{u}}) = b(\mathbf{u})\left(\bar{\mathbf{u}}\mathbf{u}^T + \mathbf{u}\bar{\mathbf{u}}^T\right) + \mathbf{u}^T\bar{\mathbf{u}}B(\mathbf{u}). \tag{A.53}$$

Next, we write out the dot product of $D'(\mathbf{u})(\bar{\mathbf{u}})\nabla c$ with $\nabla\lambda_c$ using (A.53). We note that the i,j entry in $\bar{\mathbf{u}}\mathbf{u}^T$ is defined as $\bar{\mathbf{u}}_i\mathbf{u}_j$ and the i,j entry in $\mathbf{u}\bar{\mathbf{u}}^T$ is defined as $\mathbf{u}_i\bar{\mathbf{u}}_j$.

$$
\begin{aligned}
(D'(\mathbf{u})(\bar{\mathbf{u}})\nabla c)\cdot\nabla\lambda_c &= \sum_{i=1}^{n}(D'(\mathbf{u})(\bar{\mathbf{u}})\nabla c)_i(\nabla\lambda_c)_i \\
&= \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\left(b(\mathbf{u})(\bar{\mathbf{u}}_i\mathbf{u}_j+\mathbf{u}_i\bar{\mathbf{u}}_j)+\mathbf{u}^T\bar{\mathbf{u}}(B(\mathbf{u}))_{ij}\right)(\nabla c)_j\right)(\nabla\lambda_c)_i \\
&= \sum_{i=1}^{n}\left(b(\mathbf{u})\sum_{j=1}^{n}(\bar{\mathbf{u}}_i\mathbf{u}_j+\mathbf{u}_i\bar{\mathbf{u}}_j)(\nabla c)_j+\mathbf{u}^T\bar{\mathbf{u}}\sum_{j=1}^{n}(B(\mathbf{u}))_{ij}(\nabla c)_j\right)(\nabla\lambda_c)_i \\
&= \sum_{i=1}^{n}\left(b(\mathbf{u})\bar{\mathbf{u}}_i\sum_{j=1}^{n}\mathbf{u}_j(\nabla c)_j+b(\mathbf{u})\mathbf{u}_i\sum_{j=1}^{n}\bar{\mathbf{u}}_j(\nabla c)_j+\mathbf{u}^T\bar{\mathbf{u}}(B(\mathbf{u})\nabla c)_i\right)(\nabla\lambda_c)_i \\
&= \sum_{i=1}^{n}\left(b(\mathbf{u})\bar{\mathbf{u}}_i\mathbf{u}^T\nabla c)+b(\mathbf{u})\mathbf{u}_i\bar{\mathbf{u}}^T\nabla c+\mathbf{u}^T\bar{\mathbf{u}}(B(\mathbf{u})\nabla c)_i\right)(\nabla\lambda_c)_i \\
&= b(\mathbf{u})\mathbf{u}^T\nabla c\sum_{i=1}^{n}\bar{\mathbf{u}}_i(\nabla\lambda_c)_i+b(\mathbf{u})\bar{\mathbf{u}}^T\nabla c\sum_{i=1}^{n}\mathbf{u}_i(\nabla\lambda_c)_i \\
&+ \mathbf{u}^T\bar{\mathbf{u}}\sum_{i=1}^{n}(B(\mathbf{u})\nabla c)_i(\nabla\lambda_c)_i \\
&= b(\mathbf{u})(\mathbf{u}^T\nabla c)(\bar{\mathbf{u}}^T\nabla\lambda_c)+b(\mathbf{u})(\bar{\mathbf{u}}^T\nabla c)(\mathbf{u}^T\nabla\lambda_c)+(\mathbf{u}^T\bar{\mathbf{u}})(B(\mathbf{u})\nabla c)^T\nabla\lambda_c \\
&= b(\mathbf{u})(\mathbf{u}^T\nabla c)(\nabla\lambda_c^T\bar{\mathbf{u}})+b(\mathbf{u})(\mathbf{u}^T\nabla\lambda_c)(\nabla c^T\bar{\mathbf{u}})+(B(\mathbf{u})\nabla c)^T\nabla\lambda_c(\mathbf{u}^T\bar{\mathbf{u}}) \\
&= \left(b(\mathbf{u})(\mathbf{u}^T\nabla c)\nabla\lambda_c^T+b(\mathbf{u})(\mathbf{u}^T\nabla\lambda_c)\nabla c^T+(B(\mathbf{u})\nabla c)^T\nabla\lambda_c\mathbf{u}^T\right)\bar{\mathbf{u}} \\
&= \left(\frac{\alpha_l-\alpha_t}{|\mathbf{u}|}(\mathbf{u}^T\nabla c)\nabla\lambda_c^T+\frac{\alpha_l-\alpha_t}{|\mathbf{u}|}(\mathbf{u}^T\nabla\lambda_c)\nabla c^T\right. \\
&+ \left.\left(\left(\frac{(\alpha_t-\alpha_l)E(\mathbf{u})+\alpha_t I}{|\mathbf{u}|}\right)\nabla c\right)^T\nabla\lambda_c\mathbf{u}^T\right)\bar{\mathbf{u}} \\
&= \left(\frac{\alpha_l-\alpha_t}{|\mathbf{u}|}\left((\mathbf{u}^T\nabla c)(\nabla\lambda_c^T)+(\mathbf{u}^T\nabla\lambda_c)(\nabla c^T)\right)\right. \\
&+ \left.\left(\left(\frac{(\alpha_t-\alpha_l)E(\mathbf{u})+\alpha_t I}{|\mathbf{u}|}\right)\nabla c\right)^T\nabla\lambda_c\mathbf{u}^T\right)\bar{\mathbf{u}} \\
&= \bar{\mathbf{u}}\cdot\left(\frac{\alpha_l-\alpha_t}{|\mathbf{u}|}\left(\nabla\lambda_c\nabla c^T\mathbf{u}+\nabla c\nabla\lambda_c^T\mathbf{u}\right)\right. \\
&+ \left.\mathbf{u}\nabla\lambda_c^T\left(\frac{(\alpha_t-\alpha_l)E(\mathbf{u})+\alpha_t I}{|\mathbf{u}|}\right)\nabla c\right).
\end{aligned}
$$

□

A.3 Optimal control problems

A.3.1 The Lagrangian

We want to minimize an objective function J that depends on the control u and state y. We define the optimization problem using J and a PDE constraint e.

$$\min_{(y,u)\in(Y,U)} J(y,u),$$

subject to

$$e(y,u) = 0.$$

The objective function J maps $Y \times U$ to $\mathbb{R}$ and the PDE function maps $Y \times U$ to Z, where Y, U, and Z are Banach spaces. The Lagrangian L maps $Y \times U \times Z^*$ to $\mathbb{R}$ and L is defined using the objective function and the PDE. We also introduce the variable p in Z^*.

$$L(y,u,p) = J(y,u) + < p, e(y,u) >_{Z^*,Z} .$$

We denote the duality pairing of Z and Z^* by $< \cdot, \cdot >_{Z^*,Z}$.

Theorem A.1

Riesz Representation [10]

The dual space H^* of a Hilbert space H is isometric to H itself. More precisely, for every $v \in H$, there exists a linear functional u^* defined by

$$< u^*, u >_{H^*,H} := (v,u)_H,$$

for all u in H, with u^* in H^* with norm

$$||u^*||_{H^*} = ||v||_H.$$

For any u^* in H^*, there exists a unique v in H such that

$$< u^*, u >_{H^*,H} = (v,u)_H,$$

for all u in H with v in H^* and

$$||u^*||_{H^*} = ||v||_H.$$

Using this theorem, we have that for p in Z^*, there exists a unique λ in Z such that

$$< p, e(y,u) >= (\lambda, e(y,u)).$$

From the definition of λ, we can rewrite the Lagrangian using λ instead of p.

$$L(y,u,\lambda) = J(y,u) + < p, e(y,u) >_{Z^*,Z} \tag{A.54}$$

$$= J(y,u) + (\lambda, e(y,u))_Z. \tag{A.55}$$

Frequently, notation is abused and we use p instead of λ. We now investigate further $< \lambda, e(y,u) >_{Z^*,Z}$ when e is a PDE. For simplicity, let us first assume the e is a linear PDE. We define e using A and B, which are continuous, linear operators, where A maps Y to $\mathbb{R}$ and B maps U to $\mathbb{R}$. We define f to be in $\mathbb{R}$. We can then write e using A, B, and f:

$$e(y,u) = Ay + Bu - f.$$

We can now determine explicitly the value of $< p, e(y,u) >_{Z^*,Z}$:

$$\begin{aligned} < p, e(y,u) >_{Z^*,Z} &= (\lambda, Ay + Bu - f)_Z \\ &= (\lambda, Ay)_Z + (\lambda, Bu)_Z - (\lambda, f)_Z \\ &= (A^*\lambda, y)_Y + (B^*\lambda, u)_U - (\lambda, f)_Z. \end{aligned} \tag{A.56}$$

The operators A^* and B^* are the adjoint operators of A and B, which will be defined in Section A.3.1. Let A be defined as the following PDE operator:

$$Ay = \begin{cases} \nabla \cdot (-\epsilon \nabla y + \beta y) + ry, & \text{in } \Omega, \\ 0, & \text{on } \partial\Omega. \end{cases} \tag{A.57}$$

Determining the adjoint operator

When defining the Lagrangian, recall that we had two adjoint operators A^* and B^*, which we will now define.

Definition A.1 Let $A : X \to X$ be a continuous linear operator on a Hilbert space X. Then the adjoint A^* must satisfy the following property:

$$< Ax, y >=< x, A^*y >,$$

for all x, y in X. $< \cdot, \cdot >$ is the inner product on X.

We use A defined in (A.57), using the following spaces for Y and Z:

$$Y = Z = H_0^2(\Omega).$$

We can determine the adjoint A^* and B^* using the L^2 inner product on Ω. We use integration by parts.

$$\begin{aligned}
(A^*\lambda, y)_Y = (\lambda, Ay)_Z &= \int_\Omega \lambda A y \\
&= \int_\Omega \lambda \left(\nabla \cdot (-\epsilon \nabla y + \beta y) + ry\right) \\
&= \int_\Omega \lambda \nabla \cdot (-\epsilon \nabla y) + \int_\Omega \lambda \nabla \cdot (\beta y) + \int_\Omega \lambda r y \\
&= \int_\Omega \nabla \cdot \lambda(\epsilon \nabla \cdot y) - \int_{\partial\Omega} \lambda(\nabla \epsilon y) \cdot \mathbf{n} - \int_\Omega \nabla\lambda \cdot (\beta y) \\
&\qquad + \int_{\partial\Omega} \lambda(\beta y) \cdot \mathbf{n} + \int_\Omega \lambda r y \\
&= \int_\Omega \nabla \cdot \lambda(\epsilon \nabla \cdot y) - \int_\Omega \nabla\lambda \cdot (\beta y) + \int_\Omega \lambda r y \\
&= -\int_\Omega \Delta\lambda(\epsilon y) + \int_{\partial\Omega} (\beta y)\nabla\lambda \cdot \mathbf{n} - \int_\Omega \nabla\lambda \cdot (\beta y) + \int_\Omega \lambda r y \\
&= -\int_\Omega \Delta\lambda(\epsilon y) - \int_\Omega \nabla\lambda \cdot (\beta y) + \int_\Omega \lambda r y \\
&= \int_\Omega \nabla \cdot (-\epsilon\nabla\lambda) y - \nabla\lambda \cdot (\beta y) + \lambda r y \\
&= \int_\Omega y \left(\nabla \cdot (-\epsilon \nabla\lambda) - \nabla\lambda \cdot \beta + \lambda r\right).
\end{aligned}$$

We now have the definition of the adjoint A^* in Ω.

$$A^*\lambda = \nabla \cdot (-\epsilon \nabla \lambda) - \nabla \lambda \cdot \beta + r\lambda. \tag{A.58}$$

Recall the PDE, where the operator on u is B. We assuming u lies in $L^2(\Omega)$.

$$Bu = -u.$$

Using the definition of the adjoint operator, we obtain the adjoint B^*.

$$\begin{aligned}(B^*\lambda, u)_U &= (\lambda, Bu)_Z \\ &= \int_\Omega \lambda B u \\ &= \int_\Omega \lambda(-u) \\ &= \int_\Omega (-\lambda)u.\end{aligned}$$

We now have the definition of the adjoint B^* on Ω.

$$B^*\lambda = -\lambda. \tag{A.59}$$

Combining the definition of the adjoints, (A.58) and (A.59) and definition of the Lagrangian (A.56), we have

$$(\lambda, e(y,u))_Z = (A^*\lambda, y)_Y + (B^*\lambda, u)_U - (\lambda, f)_Z, \tag{A.60}$$

$$= (\nabla \cdot (-\epsilon \nabla \lambda) - \nabla \lambda \cdot \beta + r\lambda, y)_Y - (\lambda, u)_U - (\lambda, f)_Z. \tag{A.61}$$

The derivatives of the Lagrangian

To determine the optimality conditions, we need to take the derivative of the Lagrangian with respect to each variable, and set it equal to zero. Since L is Frechet differentiable, L is also Gateaux differentiable. We use the definition of the Gateaux derivative to determine the derivatives of the Lagrangian.

Before we take the derivatives, we give explicit definitions of derivatives, taken from [10].

Let $F : U \subset X \to Y$ an operator, $X, Y, U \neq \emptyset$ open Banach spaces.

- F is *directionally differentiable* at $x \in U$ if

$$dF(x,h) = \lim_{t \to 0^+} \frac{F(x+ht) - F(x)}{t} \in Y,$$

exists $\forall h \in X$. $dF(x,h)$ is called the *directional derivative of F in the direction of h.*

- F is *Gateaux differentiable* at $x \in Y$ if F is directionally differentiable at x and the directional derivative $F'(x) : X \to Y$ is bounded and linear.

- F is *Frechet differentiable* at $x \in U$ if F is Gateaux differentiable at x and if

$$||F(x+h) - F(x) - F'(x)h||_Y = o(||h||_X),$$

for $||h||_X \to 0$.

- Chain Rule (holds for Frechet differentiable operators)

Suppose $H(x) = G(F(x))$ where F, G are Frechet differentiable at $x, F(x)$ respectively. Then H is Frechet differentiable at x with $H'(x) = G'(F(x))F'(x)$.

First, we take the derivative of the Lagrangian with respect to y:

$$
\begin{aligned}
D_y L(y,u,p)(h) &=< D_y L(y,u,p), h >_{Y^*,Y}, \\
D_y L(y,u,p)(h) &= \lim_{t\to 0^+} \frac{L(y+ht,u,p) - L(y,u,p)}{t} \\
&= \lim_{t\to 0^+} \frac{J(y+ht,u) + < p, e(y+ht,u) >_{Z^*,Z} - (J(y,u) + < p, e(y,u) >_{Z^*,Z})}{t} \\
&= \lim_{t\to 0^+} \frac{J(y+ht,u) - J(y,u)}{t} + \frac{< p, e(y+ht,u) >_{Z^*,Z} - < p, e(y,u) >_{Z^*,Z})}{t} \\
&= D_y J(y,u)(h) + \lim_{t\to 0^+} \frac{< p, e(y+ht,u) - e(y,u) >_{Z^*,Z}}{t} \\
&= D_y J(y,u)(h) + < p, \frac{\lim_{t\to 0^+} e(y+ht,u) - e(y,u)}{t} >_{Z^*,Z} \\
&= D_y J(y,u)(h) + < p, D_y e(y,u) h >_{Z^*,Z} \\
&= D_y J(y,u)(h) + < D_y e(y,u)^* p, h >_{Y^*,Y} \\
&= D_y J(y,u)(h) + D_y e(y,u)^* p(h).
\end{aligned}
$$

This gives us the derivative with respect to y:

$$D_y L(y,u,p) = D_y J(y,u) + D_y e(y,u)^* p.$$

Next, we take the derivative of the Lagrangian with respect to u:

$$D_u L(y,u,p) = D_u J(y,u) + D_u e(y,u)^* p.$$

Last, we take the derivative of the Lagrangian with respect to p:

$$
\begin{aligned}
D_pL(y,u,p)(h) &=< D_pL(y,u,p), h >_{Z^*,Z},\\
D_pL(y,u,p)(h) &= \lim_{t\to 0^+} \frac{L(y,u,p+ht) - L(y,u,p)}{t}\\
&= \lim_{t\to 0^+} \frac{J(y,u) + < p+ht, e(y,u) >_{Z^*,Z} - (J(y,u) + < p, e(y,u) >_{Z^*,Z})}{t}\\
&= \lim_{t\to 0^+} \frac{< p+ht, e(y,u) >_{Z^*,Z} - < p, e(y,u) >_{Z^*,Z})}{t}\\
&= \lim_{t\to 0^+} \frac{< ht, e(y,u) >_{Z^*,Z}}{t}\\
&=< \lim_{t\to 0^+} \frac{ht}{t}, e(y,u) >_{Z^*,Z}\\
&=< h, e(y,u) >_{Z^*,Z}\\
&= he(y,u).
\end{aligned}
$$

This gives us the derivative with respect to p:

$$D_pL(y,u,p) = e(y,u).$$

Let us define the objective function similarly to the definitions in Chapter 2-4:

$$J(y,u) = \frac{1}{2}||y-\hat{y}||^2 + \frac{\alpha}{2}||u||^2.$$

To determine the optimality conditions, we need the derivative of the objective function with respect to y and u.

First we take the derivative with respect to y:

$$\begin{aligned}
D_y J(y,u)(h) &= \lim_{t\to 0^+} \frac{J(y+ht,u)-J(y,u)}{t} \\
&= \lim_{t\to 0^+} \frac{\frac{1}{2}||y+ht-\hat{y}||^2+\frac{\alpha}{2}||u||^2-(\frac{1}{2}||y-\hat{y}||^2+\frac{\alpha}{2}||u||^2)}{t} \\
&= \lim_{t\to 0^+} \frac{\frac{1}{2}||y+ht-\hat{y}||^2-\frac{1}{2}||y-\hat{y}||^2}{t} \\
&= \lim_{t\to 0^+} \frac{\frac{1}{2}(||y+ht-\hat{y}||^2-||y-\hat{y}||^2)}{t} \\
&= \lim_{t\to 0^+} \frac{\frac{1}{2}\left((y+ht-\hat{y},y+ht-\hat{y})-(y-\hat{y},y-\hat{y})\right)}{t} \\
&= \lim_{t\to 0^+} \frac{\frac{1}{2}\left((y-\hat{y},y+ht-\hat{y})+(ht,y+ht-\hat{y})-(y-\hat{y},y-\hat{y})\right)}{t} \\
&= \lim_{t\to 0^+} \frac{\frac{1}{2}\left((y-\hat{y},ht)+(ht,y+ht-\hat{y})\right)}{t} \\
&= \lim_{t\to 0^+} \frac{\frac{1}{2}\left((ht,y-\hat{y})+(ht,y+ht-\hat{y})\right)}{t} \\
&= \lim_{t\to 0^+} \frac{\frac{1}{2}(ht,2(y-\hat{y})+ht)}{t} \\
&= \lim_{t\to 0^+} \frac{\frac{1}{2}(ht,2(y-\hat{y}))+(ht,ht)}{t} \\
&= \lim_{t\to 0^+} \frac{(y-\hat{y},ht)+(ht,ht)}{t} \\
&= \lim_{t\to 0^+} \frac{(y-\hat{y},ht)}{t}+\lim_{t\to 0^+}\frac{(ht,ht)}{t} \\
&= (y-\hat{y},\lim_{t\to 0^+} h)+,\lim_{t\to 0^+}(ht,h) \\
&= (y-\hat{y},h)+0 \\
&= (y-\hat{y},h).
\end{aligned}$$

This gives us the derivative with respect to y.

$$D_y J(y,u) = y-\hat{y}.$$

Similarly, we take the derivative of the objective function with respect to u:

$$D_u J(y,u) = \alpha u.$$

When determining the explicit form of the derivatives of the Lagrangian, we will need to use the chain rule. Given a function f that maps U to Z, where we define f below,

$$f(u) = (e(y,u),u).$$

We know that f' lies in $L(U,Z)$. Define the following functions:

$$\begin{aligned} F(u) &= (y(u),u), & F &: U \to Y \times U, \\ G(y,u) &= e(y,u), & G &: Y \times U \to Z, \\ H(u) &= G(F(u)) = e(y(u),u), & H &: U \to Z. \end{aligned}$$

From the chain rule, we know $H'(u)(h)$, but we need to determine $F'(u)(h)$ and $G'(y,u)(h_y,h_u)$.

$$H'(u)(h) = G'(F(u))F'(u)(h).$$

First, we compute $F'(u)(h)$:

$$\begin{aligned} F'(u)(h) &= \lim_{t\to 0^+} \frac{F(u+ht)-F(u)}{t} \\ &= \lim_{t\to 0^+} \frac{(y(u+ht),u+th)-(y(u),u)}{t} \\ &= \lim_{t\to 0^+} \frac{(y(u+ht)-y(u),ht)}{t} \\ &= \left(\lim_{t\to 0^+} \frac{y(u+ht)-y(u)}{t}, \lim_{t\to 0^+} h\right) \\ &= \left(y'(u)h, h\right). \end{aligned}$$

Next, we compute $G'(y,u)(h_y,h_u)$:

$$\begin{aligned} G'_y(y,u)(h) &= \lim_{t\to 0^+} \frac{G(y+ht,y)-G(y,u)}{t} \\ &= \lim_{t\to 0^+} \frac{e(y+ht,y)-e(y,u)}{t} \\ &= e'_y(y,u)h. \end{aligned}$$

We have the property that

$$G'(v,w)(h_v,h_w) = G'_v(v,w)(h_v) + G'_w(v,w)(h_w).$$

Using this property and the definition of G', we have $G'(y,u)(h_y,h_u)$:

$$\begin{aligned} G'(y,u)(h_y,h_u) &= G'_y(y,u)(h_y) + G'_u(y,u)(h_u) \\ &= e'_y(y,u)(h_y) + e'_u(y,u)(h_u). \end{aligned}$$

Combining the definition of $F'(u)(h)$ and $G'(y,u)(h_y,h_u)$, we obtain the derivative of f.

$$\begin{aligned} f'(u)(h) &= H'(u)(h) \\ &= G'(F(u))F'(u)(h) \\ &= G'(y(u),u)(y'(u)h,h) \\ &= e'_y(y,u)(y'(u)h) + e'_u(y,u)(h). \end{aligned}$$

Bibliography

[1] S. Bartels, M. Jensen, and R. Muller. Discontinuous Galerkin Finite Element Convergence for Incompressible Miscible Displacement Problems of Low Regularity. *SIAM Journal on Numerical Analysis*, 47(5):3720–2743, 2009.

[2] G. Chavent and M. Dupuy. History Matching by Use of Optimal Theory. *Society of Petroleum Engineers Journal*, 15(1), 1975.

[3] Z. Chen and R. Ewing. Mathematical Analysis for Reservoir Models. *SIAM Journal on Numerical Analysis*, 30(2):431–453, 1999.

[4] A. Dontchev, W. Hager, and V. Veliov. Second order Runge-Kutta approximations in constrained optimal control. *SIAM J. Numer. Anal.*, 38:202–226, 2000.

[5] J. Douglas, R. Ewing, and M. Wheeler. A Time Discretization Procedure for a Mixed Finite Element Approximation of Miscible Displacement in Porous Media. *RAIRO*, 17(3), 1983.

[6] Y. Epshteyn and B. Riviere. Convergence of high order methods for miscible displacement. *International Journal of Numerical Analysis and Modeling*, 5:47–63, 2008.

[7] R. Ewing and T. Russell. Efficient Time-Stepping Methods for Miscible Displacement Problems in Porous Media. *SIAM Journal on Numerical Analysis*, 19(1), February 1982.

[8] R. Ewing and M. Wheeler. Galerkin Methods for Miscible Displacement Problem in Porous Media. *SIAM Journal on Numerical Analysis*, 17(3), June 1980.

[9] X. Feng. On Existense and Uniqueness Results for a Coupled System Modeling Miscible Displacement in Porous Media. *Journal of Mathematical Analysis and Applications*, 194:883–910, 1995.

[10] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with Partial Differential Equation Constraints*, volume 23. Springer, Heidelberg, New York, Berlin, 2009.

[11] D. Leykekhman. Investigation of Commutative Properties of Discontinuous Galerkin methods in PDE Constrained Optimal Control Problems. *Journal of Scientific Computing*, 53(3):483–511, 2012.

[12] D. Leykekhman and M. Heinkenschloss. Local Error Analysis of Discontinuous Galerkin Methods for Advection-Dominated Elliptic Linear-Quadratic Optimal Control Problems. *SIAM Journal of Numerical Analysis*, 50(4):2012–2038, 2012.

[13] J. Li. Locally Mass-Conservative Method with Discontinuous Galerkin in Time for Solving Miscible Displacement Equations under Low Regularity. Master's thesis, 2013.

[14] J. Li and B. Riviere. High order discontinuous galerkin method for simulating miscible flooding in porous media. *Computational Geosciences*, To appear.

[15] J. Li, B. Riviere, and N. Walkington. Convergence of a high order method in time and space for the miscible displacement equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49:953–976, 2015.

[16] J. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*, volume 170. Springer, 1971.

[17] M. Heinkenschloss. Numerical Solution of Implicitly Constrained Optimization

Problems. *tr08-05, Dept. of Computational and Applied Mathematics, Rice University, Houston, TX*, 2008.

[18] G. Mehos and W. Ramirez. Use of Optimal Control Theory to Optimize Carbon Dioxide Miscible-Flooding Enhanced Oil Recovery. *Journal of Petroleum Science and Engineering*, 2:247–260, 1989.

[19] D. Meidner and B. Vexler. A priori error estimates for space-time finite element discretization of parabolic optimal control problems part I: Problems without control constraints. *SIAM J. Control Optim.*, 47(3):1150–1177, 2008.

[20] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006.

[21] M. Ohlberger. Convergence of a Mixed Finite Element - Finite Volume Method for the Two Phase Flow in Porous Media. *East-West Journal of Numerical Mathematics*, 5:183–210, 1997.

[22] B. Riviere and N. Walkington. Convergence of Discontinuous Galerkin Method forthe Miscible Displacement Equation Under Low Regularity. *SIAM Journal on Numerical Analysis*, 49(3):1085–1110, 2011.

[23] B. Riviere and M. Wheeler. Discontinuous Galerkin methods for flow and transport problems in porous media. *Communications in Numerical Methods in Engineering*, 18:63–68, 2002.

[24] T. Russell. Time Stepping Along Characteristics with Incomplete Iteration for a Galerkin Approximation of Miscible Displacement Problems in Porous Media. *SIAM Journal on Numerical Analysis*, 22(5), October 1985.

[25] M. Simon and M. Ulbrich. Optimal control of partially miscible two-phase flow with applications to subsurface CO_2 sequestration. In M. Bader, H.-J. Bungartz,

and T. Weinzierl, editors, *Advanced Computing*, volume 93 of *Lecture Notes in Computational Science and Engineering*, pages 81–98. Springer Berlin Heidelberg, 2013.

[26] M. Simon and M. Ulbrich. Adjoint based optimal control of partially miscible two-phase flow in porous media with applications to CO_2 sequestration in underground reservoirs. *Optimization and Engineering*, 16(1):103–130, 2015.

[27] S. Sun, B. Riviere, and M. Wheeler. A combined mixed finite element method and discontinous Galerkin method for miscible displacement problem in porous media. *Recent Progress in Computational and Applied PDEs, Kluwer/Plenum, New York*, pages 323–348, 2002.

[28] F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, volume 112. American Mathematical Society, 2010.

[29] H. Yücel, M. Heinkenschloss, and B. Karasözen. Distributed Optimal Control of Diffusion-Convection-Reaction Equations Using Discontinuous Galerkin Methods. *Numerical Mathematics and Advanced Applications 2011. Springer Berlin Heidelberg*, pages 389–397, 2011.

[30] D. Zeitoun and G. Pinder. An optimal control least squares method for solving coupled flow-transport systems. *Water Resources Research*, 29:217–227, 1993.